PhD Thesis/ Tesis Doctoral

Multilingual Data Collection for Multiple Corpus-based Approaches to Translation and Interpretation

Creación de Datos Multilingües para Diversos Enfoques Basados en Corpus en el Ámbito de la Traducción y la Interpretación

Hernani Pereira Gomes da Costa

Supervisors/directores:
Dra. D.ª Gloria Corpas Pastor
Dr. D. Ruslan Mitkov
Dra. D.ª Miriam Seghiri Domínguez

Programa de Doctorado en Lingüística, Literatura y Traducción
Faculty of Philosophy and Humanities/ Facultad de Filosofía y Letras
University of Malaga/ Universidad de Málaga
2019

AUTOR: Hernani Pereira Gomes Da Costa

http://orcid.org/0000-0002-6813-4641

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

UNIVERSIDAD DE MALAGA

DEPARTAMENTO DE TRADUCCIÓN E INTERPRETACIÓN

A quien corresponda:

Por la presente, en calidad de directores y tutor de la tesis doctoral de D. **Hernani Pereira Gomes Da Costa**, hacemos constar que el trabajo depositado cumple con todos los criterios exigidos en el Programa de Doctorado en el cual se enmarca. El doctorando presenta una metodología y resultados objetivos y realistas, a la par que cuenta con unas claras y detalladas conclusiones, seguidas de una bibliografía extensa y actualizada.

Asimismo, confirmamos que las publicaciones que avalan el trabajo no han sido utilizadas en tesis anteriores y que ninguno de los miembros propuestos para formar parte del tribunal o evaluadores externos figura como co-autor en ninguna de las publicaciones que se incluyen.

Por todo lo mencionado anteriormente, consideramos que la presente tesis doctoral cumple de manera holgada con todos los requisitos exigibles para su pública defensa y expresamos nuestro visto bueno para la admisión a trámite de la misma.

Y para que surtan los efectos oportunos, lo firmamos en Málaga, a 27 de noviembre de 2018.

Atentamente,

David Marín	Míriam Seghiri	Gloria Corpas	Ruslan Mitkov
Tutor	Directora	Directora	Director

Table of Contents

Acronyms

CL: Computational Linguistics

DSM: Distribution Similarity Measures

IE: Information Extraction

IR: Information Retrieval

ML: Machine Learning

MT: Machine Translation

NCT: Number of Common Tokens

NER: Named Entity Recognition

NLP: Natural Language Processing

POS: Part-of-speech

SCC: Spearman's Rank Correlation Coefficient

TM: Translation Memories

TMT: Terminology Management Tools

TET: Terminology Extraction Tools

List of Figures

List of Tables

Declaration

"Real difficulties can be overcome;
it is only the imaginary ones that are
unconquerable."

—Theodore N. Vail

This thesis was submitted for the degree of Doctor of Philosophy at the University of Malaga, Spain. The research described herein was conducted under the supervision of Prof Gloria Corpas Pastor, Prof Ruslan Mitkov and Dr Miriam Seghiri in the Department of Translation and Interpreting at the Faculty of Philosophy and Humanities, between 2013 and 2018.

This work is to the best of my knowledge original, except when acknowledgements and references are made to previous work. Neither this, nor any substantially similar dissertation has been or is being submitted for any other degree, diploma or other qualification at any other university as far as I am aware.

Acknowledgements

"You are the average of the five people you spend the most time with."

—Jim Rohn

From the first paper read to finally write the last page of my thesis, this PhD has been a long and sometimes winding road. Along the way, I have been fortunate to have the support and guidance of many people to whom I wish to express my heartfelt gratitude.

I would like to start by thanking my main supervisor Prof Gloria Corpas Pastor for her enthusiasm, guidance, wisdom and endless support. I am also extremely grateful to my other two supervisors, Prof Ruslan Mitkov and Dr Miriam Seghiri, who where always there for listening me, for giving me their expert opinion on every detail of this work, for encourage me to go further, for their endless support, and most importantly for their friendship.

I own my deepest gratitude to my colleagues and friends Hanna Béchara, Dr Shiva Taslimipoor, Dr Rohit Gupta, Dr Marcos Zampieri and specially Dr Isabel Durán Muñoz for the long and interesting discussions that resulted very fruitful in a form of various scientific publications, and for believing in my ideas.

I would like to thank everybody in the LEXYTRAD research group and the people from the Department of Translation and Interpreting because they were somehow important for the completion of this work. Special remarks to Ruth Guitiérrez Florido, with whom I worked together on the organisation of various research events and always helped me solving all kinds of university-related paperwork.

I would also like to thank João Franco and Bruno Antunes for proofreading this thesis.

My PhD work was financially supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n. 317471. The research reported in this thesis has also been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015), the Educational Innovation Project NOVATIC (PIE 15-145, 2015-2017), the INTERPRETA 2.0 (PIE17-015, 2017-2019), the R&D project INTELITERM (ref. n. FFI2012-38881, 2012-2015), the R&D LATEST (ref. n. 327197-FP7-PEOPLE-2012-IEF) project, the R&D Project for Excellence TERMITUR (ref. n. HUM2754, 2014-2017), the VIP R&D Project

(ref. n. FFI2016-75831), the TACTRAD teaching network and TRAJUTEC thematic network (University of Malaga).

Apart from the financial support, I am also indebted to the Marie Currie Program for the opportunity to meet and learn from extraordinary professionals that I encountered during the last couple of years from all over the world during my secondments, conferences and meetings I had the opportunity to attend. I would like to thank all the Early Stage Researches (ESRs) and Experienced Researches (ERs) involved in the EXPERT (EXPloiting Empirical appRoaches to Translation) project for their constant support, help and companion during this journey. A special thanks to my friends and to all the people I met during my secondments at the University of Wolverhampton and Translated s.r.l. Moreover, I also need to say a big thank you to all the supervisors, researchers and administrative staff involved in the project. Without them, the project would not have been such a success. A special thanks to Dr Constantin Orasan, who along his team organised and managed the EXPERT project, created a research network with strong collaboration and exchange of ideas and made it a fruitful source of high-quality research and a great experience for all its participants. For this and so much more, thank you Marie Curie and all those involved in the EXPERT consortium.

I would like to give a special word to my friends Dr Bruno Antunes, Dr João Franco and Dr Hugo Gonçalves Oliveira that always believed in me and always inspired me to pursue my dreams.

I would like to express my deepest gratitude to my PhD fellow and friend Dr Anna Zaretskaya for those endless hours of inspiring conversations and invaluable peer support.

Finally, I would like to take this opportunity to express my gratitude to my family for their encouragement and support –my parents, Hernani Costa and Silvina Costa and my brother, Paulo Costa. A special thanks to my beloved wife Priscila Costa and our precious daughters Chloe Costa and Amy Costa for their endless love, patience, unfailing encouragement, support and understanding during these years. Thank you for being by my side every step of the way.

Hernani Costa
November 2018

Preface

"If you can't explain it to a six year old,
you don't understand it yourself"

—Albert Einstein

About seven years ago, almost by accident, I ended up engaging in an academic research career. It all started in 2008, during my European Union student exchange programme (Erasmus), at the University of Vigo[1], when I enrolled an introductory class in Natural Language Processing (NLP), which made me wonder how could machines understand and process the human language. Then, one year later I had to make an important decision, either do my Master's dissertation in a company or in the university within a research group. This simple choice changed my entire professional life.

After deciding to do my Master's dissertation in the Department of Informatics Engineering[2], I knew that it should be somehow related with NLP. That is why I decided to join the Onto.PT[3], an ambitious research project lead by Dr Hugo Gonçalves Oliveira and Dr Paulo Gomes that aimed at building the first lexical ontology for Portuguese. Once I started to understand more about the subject and the current challenges, I realised that I was spending most of my days reading and exploring new algorithms to extract and validate semantic knowledge from text – it turn to be a really exciting topic for me. Even without noticing I finished my Master's dissertation as well as my Master in Computer Science in September 2010. Two months later, in November 2010, I started working as a full-time researcher in a different project at the same research center. During the next three years I not only had the opportunity to expand my comfort zone, but also to work as a NLP researcher for Linguateca[4] and as a lecturer for two different institutions.

Sooner my dreams got bigger and I decided to go abroad and pursuit a PhD degree. That is when I came across this opportunity to be part of a this huge Machine Translation (MT) project named EXPERT (EXPloiting Empirical appRoaches to Translation), which involved six European universities (University

[1] `https://www.uvigo.gal`
[2] `http://www.dei.uc.pt`
[3] `http://ontopt.dei.uc.pt`
[4] `https://www.linguateca.pt`

of Wolverhampton[5], University of Malaga[6], University of Sheffield[7], University of Amsterdam[8], University of Saarland[9] and Dublin City University[10]) and five international leading language service provider companies (Pangeanic[11], Translated s.r.l.[12], Hermes Traducciones y Servicios Lingüísticos[13], Wordfast[14] and Etrad[15]). I never had work in this area before, but I knew that it would have something in common with what I had been doing so far –after all, broadly speaking, we can say that MT is also about making machines understand the human language. Although I would not have the opportunity to work with my mother tongue (Portuguese), working with English and Spanish turned to be more challenging but at same time rewarding.

Working as a Early Stage Researcher (ESR) in the EXPERT project turned to be very important for improving and developing my hard and soft skills as a Researcher as well as a person. Sincerely, I feel that during this journey I contributed with various interesting scientific publications and open source NLP tools that helpfully will be used and improved by the scientific community. To conclude, I would like to to say that this entire experience resulted to be so amazing and enriching that I will always cherish and treasure forever.

[5] https://www.wlv.ac.uk
[6] https://www.uma.es
[7] https://www.sheffield.ac.uk
[8] http://www.illc.uva.nl
[9] https://www.uni-saarland.de
[10] https://www.dcu.ie
[11] http://www.pangeanic.com
[12] http://www.translated.net
[13] http://www.hermestrans.com
[14] http://www.wordfast.com
[15] http://www.etrad.com.ar

List of Original Contributions

"Do what you think is interesting,
do something that you think is fun and worthwhile,
because otherwise you won't do it well anyway."

—Brian W. Kernighan

This research is supported by the following peer-reviewed scientific publications (listed in a chronological order). To access the full content of them, please see Appendix A.

- Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014). Technology-assisted Interpreting. *MultiLingual* #143, 25(3):27-32.

- Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014). A comparative User Evaluation of Terminology Management Tools for Interpreters. In *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, pages 68-76, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

- Costa, H., Corpas Pastor, G., and Seghiri, M. (2014). iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, London, UK.

- Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015). iCorpora: Compiling, Managing and Exploring Multilingual Data. In *7th Int. Conf. of the Iberian Association of Translation and Interpreting Studies, AIETI*, pages 74-76, Malaga, Spain.

- Costa, H., Béchara, H., Taslimipoor, S., Gupta, R., Orasan, C., Corpas Pastor, G., and Mitkov, R. (2015). MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. In *9th Int. Workshop on Semantic Evaluation, SemEval'15*, pages 96-101, Denver, Colorado. ACL.

- Costa, H. (2015). Assessing Comparable Corpora through Distributional Similarity Measures. In *EXPERT Scientific and Technological Workshop*, pages 23-32, Malaga, Spain. Tradulex.

- Zampieri, M., Gebrekidan Gebre, B., Costa, H., and van Genabith, J. (2015). Comparing Approaches to the Identification of Similar Languages. In *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial'15), 2nd Discriminating between Similar Languages Shared Task (DSL'15)*, pages 7, Hissar, Bulgaria.

- Costa, H., Corpas Pastor, G., and Mitkov, R. (2015). Measuring the Relatedness between Documents in Comparable Corpora. In *11th Int. Conf. on Terminology and Artificial Intelligence, TIA'15*, pages 29-37, Granada, Spain.

- Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2015). An Interpreters' Guide to Selecting Terminology Management Tools. In *NATO Conf. on Terminology Management*, Brussels, Belgium.

- Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015). Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora. In *New Horizons in Translation and Interpreting Studies (Full papers)*, pages 133-141, Geneva, Switzerland. Tradulex.

- Costa, H., Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2016). Nine terminology extraction Tools: Are they useful for translators? *MultiLingual* #159, 27(3).

- Costa, H., Durán Muñoz, I., Corpas Pastor, G., and Mitkov, R. (2016b). Compilação de Corpos Comparáveis Especializados: Devemos sempre confiar nas Ferramentas de Compilação Semi-automáticas? *Linguamática*, 7(2):17.

- Costa, H. and Corpas Pastor, G. and Durán Muñoz, I. (2017) Assessing Terminology Management Systems for Interpreters. In Corpas Pastor, Gloria and Durán Muñoz, Isabel, *Trends in E-Tools and Resources for Translators and Interpreters*, volume 45, pages 57-84, Brill.

Apart from the aforementioned original contributions, various computer programs and tools were developed and made publicly available for public use. Hereafter, we present them and summarise their main functionalities.

- iCompileCorpora: a web interface that guides the user through the creation of mono- and multilingual comparable corpora;

- SCleaner: a web application that helps users to format text copied from a PDF file;

- PreProcessor: a program that helps users to annotate raw textual data;

- STSModule: a program that aims to help users computing the semantic similarity between sentences and documents in English;

- DSMModule: a program that helps the user to assess and rank comparable documents according to their internal degree of similarity.

It is important to mention that all the software was made publicly available and, thus free for being used and edited by anyone, both in a research and in a commercial setting. We believe this is the best way to contribute for the advancement of science in general and Computational Linguistics (CL) in particular.

Abstract

"Begin at the beginning,"
–the King said, gravely–
"and go on till you come to an end;
then stop".

—Lewis Carroll, *Alice in Wonderland*

Corpora are playing an increasingly important role in our multilingual society. High-quality parallel corpora are a preferred resource in the language engineering and the linguistics communities. Nevertheless, the lack of sufficient and up-to-date parallel corpora, especially for narrow domains and poorly-resourced languages is currently one of the major obstacles to further advancement across various areas like translation, language learning and, automatic and assisted translation. An alternative is the use of comparable corpora, which are easier and faster to compile. Corpora, in general, are extremely important for tasks like translation, extraction, inter-linguistic comparisons and discoveries or even to lexicographical resources. Its objectivity, reusability, multiplicity and applicability of uses, easy handling and quick access to large volume of data are just an example of their advantages over other types of limited resources like thesauri or dictionaries. By a way of example, new terms are coined on a daily basis and dictionaries cannot keep up with the rate of emergence of new terms.

Accordingly, this research work aims at exploiting and developing new technologies and methods to better ascertain not only translators' and interpreters' needs, but also professionals' and ordinary people's on their daily tasks, such as corpora and terminology compilation and management. The main topics covered by this work relate to Computational Linguistics (CL), Natural Language Processing (NLP), Machine Translation (MT), Comparable Corpora, Distributional Similarity Measures (DSM), Terminology Extraction Tools (TET) and Terminology Management Tools (TMT). In particular, this work examines three main questions: 1) Is it possible to create a simpler and user-friendly comparable corpora compilation tool? 2) How to identify the most suitable TMT and TET for a given translation or interpreting task? 3) How to automatically assess and measure the internal degree of relatedness in comparable corpora? This work is composed of thirteen peer-reviewed scientific publications, which are included in the Appendix A, while the methodology used and the results obtained in these studies are summarised in the main body of this document.

The first task was approached by doing an extensive analysis on the

existing comparable compilation tools on the market, which their limitations and strengths were reported and considered while a new multilingual comparable corpora prototype, named iComparableCorpora was created. iComparableCorpora aimed not only to overcome various spotted usability problems, limitations and performance issues, but also to improve the compilation process flexibility and robustness.

The second task of this research focused on addressing translators' and interpreters' needs and suggest new methodologies or tools to help them increase the productivity and ease their labour-intensive activities. To do so, a set of users' requirements was careful compiled from various users' surveys in the literature. In parallel, a set of features offered by the most known TMT and TET on the market was also identified. Then, by matching the software functionalities offered by these tools with the users' requirements, two new standardised methodologies capable of evaluating the current TMT and TET on the market were proposed. Finally, new directions of improvement were also suggest mostly do to the current displacement between the users' needs and offered software functionalities.

The third and last research task of this research mainly focused on exploring various methods capable of helping users accessing comparable corpora. In detail, a simple, yet efficient methodology capable of assessing and ranking comparable documents according to their internal degree of similarity was proposed. This method not only can help the user to have a better idea about the quality of the documents in the corpus but also can help deciding which documents should belong or be removed from it.

Along this journey, various programs and tools were created. Two of them resulted from the first research question. Namely SCleaner, a web application that helps users to format text copied from a PDF file, and iCompileCorpora, a web interface that guides the user through the creation of multilingual comparable corpora. Regarding the third research question, three programs were created: PreProcessor, a program that helps users to annotate raw textual data; STSModule, a program that aims at helping users computing the semantic similarity between sentences and documents in English; and, finally DSMModule, a program that helps the user to assess and rank documents according to their internal degree of similarity.

Keywords: comparable corpora, computational linguistics, distributional similarity measures, human translation, interpretation, machine translation, natural language processing, terminology extraction tools, terminology management tools.

Resumen

"Empieza por el principio"
–dijo el Rey con gravedad–
"y sigue hasta llegar al final;
allí te paras".

—Lewis Carroll, *Alice in Wonderland*

En la actual sociedad multilingüe, los corpus lingüísticos desempeñan, a día de hoy, un papel cada vez más importante. Entre sus principales ventajas destaca que los corpus lingüísticos son de gran utilidad en el desempeño de tareas traductológicas, de extracción y análisis terminológico, de comparaciones interlingüísticas, así como para los estudios lexicográficos (Aston, 2016; Gil-Berrozpe and Faber, 2016). Asimismo, su casi inmediata accesibilidad a un gran volumen de datos a la par que su objetividad, reutilización y versatilidad son algunas de las ventajas de los corpus frente a otro tipo de recursos más limitados, tales como tesauros o diccionarios que, en ocasiones, son incapaces de mantenerse actualizados a la acuñación de términos que se produce casi a diario (Mitkov, 2016).

Concretamente, los corpus paralelos, es decir, aquella recopilación de textos en el que cada uno de ellos se traduce a uno o más idiomas distintos del original (EAGLES, 1996b), se están perfilando como el recurso preferido en campos como la ingeniería lingüística –fundamentalmente, para el procesamiento del lenguaje natural o PLN (Jurafsky and Martin, 2009)– o en los estudios lingüísticos, en general (Cencini, 2002; Kotani and Yoshimi, 2015; Laviosa, 2016). Sin embargo, la carencia de corpus ya existentes en discursos con un grado de especialización constituye uno de los mayores retos en el desarrollo de disciplinas como la traductología, en general –especialmente, de la traducción automática y asistida (Poibeau, 2017)– o el aprendizaje de idiomas (Meunier and Dymetman, 2014).

De esta forma, recurrir a textos bilingües y multilingües, no paralelos, también conocidos como corpus comparables –es decir, integrado por muestras textuales similares originales en uno o más idiomas que utilizan los mismos criterios de diseño (EAGLES, 1996b; Corpas Pastor, 2001:158; Maia, 2003)– constituiría un enfoque alternativo debido a su más rápida y sencilla compilación. De esta forma, los corpus comparables, ya sean "fuertemente comparables" o "débilmente comparables" (Skadiņa et al., 2010a), han sido ampliamente empleados como recursos en traducción automática (Rapp et al., 2016), traducción profesional (Corpas Pastor and Seghiri, 2009; Seghiri, 2015; 2017b; Arce Romeral and Seghiri, 2018b) o interpretación, tanto en el desarrollo del ámbito profesional como en

el científico (Cencini, 2002; Straniero S., 2012; Fantinuoli and Zanettin, 2015; Defrancq, 2016; Corpas Pastor and Seghiri, 2016; Seghiri, 2017a; Arce Romeral and Seghiri, 2018a; Pérez-Pérez, 2018). Su efectividad ha sido, asimismo, difundida en iniciativas como *Building and Using Comparable Corpora* (BUCC[16]), que comparte los resultados de sus investigaciones en cuanto a la versatilidad de los corpus comparables en los de estudios e investigaciones multilingües desde el año 2007.

A pesar de que los corpus comparables han sido capaces de compensar la escasez de recursos lingüísticos y, en última instancia, mejorar la calidad de las traducciones automáticas, especialmente para idiomas con escasos recursos y discursos altamente especializados (Munteanu and Marcu, 2005; Eisele and Xu, 2010; Skadiņa et al., 2010a), tal y como afirma la autora Maia, 2003, la recolección de dichos datos supone un significativo desafío. En la actualidad, se puede abordar el proceso de compilación de corpus manualmente e incluso recurrir a herramientas especializadas diseñadas para automatizar dicha tarea (Baroni and Bernardini, 2004; Baroni et al., 2006; de Groc, 2011). Sin embargo, estos recursos de compilación presentan algunas dificultades ya que, o bien son muy escasos o las funcionalidades que ofrecen son muy limitadas (Gutiérrez Florido et al., 2013; Costa et al., 2014c); en definitiva, su simplicidad redunda negativamente en su usabilidad y rendimiento. A modo de ejemplo, no permiten la recopilación de más de un corpus comparable al mismo tiempo ni el uso de más de un operador booleano cuando se crean cadenas de búsqueda (Baroni and Bernardini, 2004; Baroni et al., 2006; de Groc, 2011). De hecho, no solo en fase de compilación, sino también en la posterior, la fase de explotación, las herramientas de gestión de corpus actuales no satisfacen las necesidades profesionales de los usuarios que las emplean.

Las limitaciones señaladas se traducen en la necesidad acuciante, tanto en la mejora como en el diseño de nuevas herramientas de compilación adaptadas a las necesidades de los traductores e intérpretes (Costa et al., 2014c; 2015d;e), tal y como han demostrado los resultados de recientes encuestas científicas (cf. Rodríguez and Schnell, 2009; Bilgen, 2011; Durán Muñoz, 2012; Zaretskaya et al., 2015) sobre la disposición de los profesionales en traducción e interpretación ante las herramientas tecnológicas disponibles en el mercado. Concretamente, los trabajos de referencia pusieron de relieve la importancia de investigar con mayor exhaustividad las razones que dificultan a la gran mayoría de los traductores e intérpretes la utilización de recursos tecnológicos, así como la necesidad de mejorar o implementar nuevas herramientas y metodologías capaces de ayudar a los profesionales lingüísticos a automatizar algunas de sus tareas, tales como gestión y extracción terminológica, ya sea en un entorno monolingüe o multilingüe.

Otro de los mayores escollos en torno al concepto de corpus comparable está relacionado con la forma y el contenido de las muestras textuales que lo integran, aspectos de suma importancia en la recopilación de los documentos y la obtención de óptimos resultados dimanantes de su análisis. Así, el *Expert Advisory Group on Language Engineering Standards Guidelines* (EAGLES, 1996b) definió el concepto de corpus comparable de la siguiente manera: "A comparable corpus is one which selects similar texts in more than one language or variety. There is as yet no agreement on the nature of the similarity, because there are very few examples of comparable corpora", afirmación de la cual se extrae que no existe consenso en cuanto a su similitud de un corpus (Braschler and Scäuble, 1998; Maia, 2003;

[16] `https://comparable.limsi.fr/bucc2016/`

Bekavac et al., 2004; Fung and Cheung, 2004; Skadiņa et al., 2010a). Aunque sí se han llevado a cabo incipientes trabajos para estudiar la determinación del grado de comparabilidad de los corpus comparables (Skadiņa et al., 2010b:162), aún no se dispone de una metodología unificada y estandarizada para describir, evaluar y clasificar automáticamente los documentos de acuerdo a su grado interrelación.

De manera general, si bien es cierto que en el mercado de la traducción existe una gran cantidad de herramientas informáticas, los intérpretes tampoco se han beneficiado de este desarrollo y avance tecnológico (Durán Muñoz, 2012; Costa et al., 2014b; 2016b; Corpas Pastor and Durán Muñoz, 2017; Corpas Pastor, 2018), lo que se traduce en una urgente necesidad por mejorar la tecnología actual, especialmente en el ámbito de la interpretación.

En definitiva y a pesar de las mejoras que aún se deben afrontar, los corpus comparables ocupan un papel muy importante en el desarrollo de la actividad profesional diaria de los traductores e intérpretes. En consecuencia, el presente trabajo de investigación tiene como objetivo explotar y desarrollar nuevas tecnologías y métodos para definir con mayor decisión, no solo las necesidades de los traductores e intérpretes, sino también la de los profesionales y personas no especializadas en la materia en tareas como la compilación y explotación de corpus y terminología. Por las razones anteriormente expuestas, las principales líneas de investigación abordadas en este trabajo están relacionadas con la lingüística computacional (*Computacional Linguistics*, CL); el procesamiento del lenguaje natural, PLN (*Natural Language Processing*, NLP); la traducción automática (*Machine Translation*, MT); corpus comparables, interpretación, traducción humana, medidas de similitud distributiva (*Distributional Similarity Measures*, DSM); herramientas de extracción terminológica (*Terminolgy Extraction Tools*, TET) y las herramientas de gestión de terminológica (*Terminology Management Tools*, TMT). Aunque los temas se exponen con más detalle a continuación, la investigación de este trabajo gira entorno a las siguientes cuestiones principales (a las que se le ha llamado *Research Question*, RQ):

1) ¿Es posible la creación de una herramienta de compilación de corpus comparables de manejo más sencillo?

2) ¿Cómo se pueden identificar las herramientas de gestión y extracción terminológica más adecuadas para una determinada traducción o interpretación?

3) ¿Cómo evaluar y medir automáticamente el grado de interrelación de los corpus comparables?

Asimismo, el presente estudio se compone de trece publicaciones científicas revisadas por pares incluidas en el Apéndice A (cfr. Appendix A), Asimismo, la terminología utilizada y los resultados obtenidos en el presente estudio se exponen a lo largo del desarrollo de este trabajo.

En este contexto de importancia creciente de demanda de las herramientas multilingües, el primer objetivo de esta tesis es doble. En primer lugar, se analizan las herramientas de compilación de corpus más conocidas en el mercado mediante la identificación de sus limitaciones y la consecuente propuesta de mejoras. Seguidamente, y tomando como referencia el estudio previo, se ha diseñado y desarrollado un prototipo novedoso, flexible y fiable basado en la web,

capaz de explotar corpus virtuales mono- y multilingües, al que hemos llamado *iComparableCorpora*[17,18], cuyas características se detallarán a continuación.

La segunda parte de esta investigación se ha centrado, a partir de los hallazgos científicos publicados en encuestas realizadas, en el análisis de las necesidades de los traductores e intérpretes, así como en la propuesta de nuevas tecnologías o herramientas que redunden en beneficio de la productividad y gestión del tiempo de dichos profesionales, gracias a la previa identificación de las funcionalidades que ofrecen las herramientas de gestión y extracción terminológica más populares en el merado profesional.

En cuanto a su grado de comparabilidad, los corpus comparables (al estar integrados por textos originales, y no de sus traducciones) pueden oscilar desde los poco a los altamente comparable. Si bien, el concepto de comparabilidad apenas ha sido abordado científicamente o, si ha sido objeto de estudio, no existe consenso en torno a su este. De esta manera, el tercer objetivo principal de este trabajo consiste en la exploración de nuevos métodos capaces de facilitar el acceso a los corpus comparables capaces de filtrar, de manera automática, los documentos irrelevantes y que, por tanto, mejoren la calidad del corpus. De hecho, uno de los inconvenientes más importantes cuando hablamos de compilación automática es el ruido documental (Costa et al., 2015c), ya que los investigadores se ven obligados a realizar una supervisión estricta a posteriori para reducirlo y que evite, en consecuencia, posibles problemas durante su análisis posterior. En este sentido, el presente trabajo apunta a la implementación de sencillas metodologías, pero eficientes, capaces de evaluar y clasificar los documentos de acuerdo a su grado de similitud. Para ello, las medidas de similitud distributivas se combinan con técnicas de PLN para evaluar el grado de relación interna del corpus. En definitiva, la metodología resultante permitirá no solo agrupar los documentos, sino también extraer información sobre el corpus en cuestión. Como resultado de este tercer objetivo de investigación, se han creado tres programas informáticos: *PreProcessor*[19], que ayuda a los usuarios a anotar datos textuales sin procesar; *STSModule*[20], *Semantic Textual Similarity Module* (módulo de similitud textual semántica), que surge con el objetivo de calcular la similitud semántica entre oraciones y documentos en inglés; y, finalmente, *DSMModule*[21], *Distributional Similarity Measures Module* (módulo de medidas de similitud distributiva), un programa que permite evaluar y agrupar los documentos según su grado interno de similitud. Las características y funcionalidades de los programa inmediatamente mencionados se detallarán a lo largo del presente apartado.

Una vez establecidos los objetivos de nuestra investigación, para acercarnos a estos, nos hemos propuesto las siguientes cuestiones:

> **RQ1**: ¿Es posible crear una herramienta de compilación de corpus comparables de manejo más sencillo?
>
> a) ¿Es factible permitir que el usuario compile más de un corpus multilingüe comparable al mismo tiempo en lugar de un solo corpus monolingüe?

[17] `https://github.com/hpcosta/iCompileCorpora`
[18] `https://icompilecorpora.herokuapp.com/home`
[19] `https://github.com/hpcosta/PreProcessor`
[20] `https://github.com/hpcosta/STSModule`
[21] `https://github.com/hpcosta/DSMModule`

b) ¿Podemos resolver algunos problemas de usabilidad de las herramientas de compilación actuales?

c) ¿Cómo simplificar el proceso de compilación para satisfacer no solo las necesidades de los traductores e intérpretes, sino también las necesidades de otros profesionales y personas no especializadas en la materia?

RQ2: ¿Cómo identificar las herramientas de extracción y gestión terminológica más adecuadas para una tarea de traducción o interpretación determinada?

a) ¿Es posible identificar las funcionalidades de las herramientas de extracción y gestión terminológica más requeridas por los usuarios que las emplean?

b) ¿Cómo se podrían trasladar las características previamente detectadas a un sistema de evaluación estandarizado?

c) ¿Cuáles pueden ser las mejoras que podrían implementarse en las herramientas de gestión y extracción terminológica actuales para satisfacer las necesidades profesionales de los traductores e intérpretes?

RQ3: ¿Cómo evaluar y medir automáticamente el grado relación entre los corpus comparables?

a) ¿Es posible evaluar automáticamente el grado interno de comparabilidad entre oraciones, documentos o incluso entre cuerpos?

b) ¿Cómo pueden combinarse los métodos de PLN y estadísticos para construir métodos automáticos capaces de evaluar y clasificar oraciones y documentos de acuerdo con su grado de comparabilidad?

c) ¿Se puede mejorar la calidad interna de los corpus comparables al filtrar documentos con un bajo grado de relación?

En cuanto a la estructura, la investigación se ha organizado en seis capítulos: El primer capítulo presenta el contexto de investigación, problemas, objetivos y enfoques, así como las principales contribuciones. El Capítulo 2 está dedicado al marco teórico de la investigación. En primer lugar, se aborda el concepto de corpus y sus fases de diseño y protocolo de compilación para, seguidamente, proponer nuevos métodos que evalúen el grado interno de compatibilidad de los corpus comparables. A continuación, el Capítulo 3 se dedica a los diferentes aspectos de la primera RQ. Así, este capítulo tiene como objetivo el desarrollo de una aplicación web sencilla, a la que hemos llamado *iCompileCorpora*, diseñada para la compilación de corpus comparables multilingües, previo análisis de las deficiencias y fortalezas de las herramientas de compilación de corpus comparables más conocidas en el mercado. Finalmente, se han propuesto algunas ideas de mejora para abordar futuras investigaciones. El Capítulo 4 explora y propone varios métodos para evaluar las herramientas terminológicas en el ámbito profesional de la traducción e interpretación. Así, con el propósito de afrontar la segunda RQ, este capítulo comienza con la exposición de un listado de las herramientas que asisten al desarrollo profesional de la interpretación. A continuación, se estudian las características

que los intérpretes esperan de una de herramienta de gestión de terminología y se propone un sistema estandarizado para la evaluación de las ya existentes en el mercado. En tercer lugar, y de igual manera, se analizan las funcionalidades que debe reunir una herramienta de gestión de terminología y se realiza una comparación de las más conocidas. Finalmente, en la última sección, se presentan los principales hallazgos en torno a una posible actualización de las herramientas de terminología utilizadas por traductores e intérpretes. El Capítulo 5, previa presentación del marco teórico, se ilustra una metodología para evaluar y clasificar automáticamente los documentos, de acuerdo a su grado interno de relación de los corpus comprables a la par que se detallan las diversas técnicas de PLN y sistemas estadísticos involucrados. Posteriormente, se discuten los resultados obtenidos y se sugieren futuras líneas de investigación. El capítulo se cierra con una conclusión final del trabajo de investigación y expone sus principales contribuciones. El trabajo finaliza con el Apéndice A que reproduce, por orden cronológico, las publicaciones resultantes de todo el trabajo de investigación abordado.

Tal y como se ha adelantado previamente, a lo largo de esta investigación se crearon diversos programas como resultado de las RQ establecidas al inicio, cuyas características se detallan a continuación:

Por lo que se refiere a la primera RQ se han implementado dos herramientas, a saber, *SCleaner*[22,23] y *iCompileCorpora*. En lo que respecta a *SCleaner*, se trata de un programa basado en la web que ayuda a los usuarios a dar formato a un texto copiado de un archivo en formato .pdf. A modo de ejemplo, elimina las tabulaciones y espacios en blanco adicionales y divide las oraciones automáticamente en la forma apropiada. Por su parte, *iCompileCopora* es una interfaz web que guía al usuario a través de la creación de un corpus virtual. Diseñado tanto para principiantes como para expertos en el campo, *iCompileCorpora* no solo se presenta en un formato sencillo con pasos simplificados, sino que también permite a los usuarios avanzados establecer opciones de compilación avanzadas durante el proceso. En definitiva, se trata de un recurso cuyo objetivo es aumentar la flexibilidad y fiabilidad del proceso de compilación. En ese sentido, es importante puntualizar que la finalidad no es la creación de una herramienta comercial, sino un concepto de prueba que establezca las bases y la dirección inicial de cara a su futuro avance y desarrollo. En definitiva, se trata de un recurso con un funcionamiento muy intuitivo que permite al usuario opciones como la compilación de más de un corpus comparable al mismo tiempo o el uso de varios operadores booleanos en la creación de consulta de búsqueda, entre otras funcionalidades relacionadas con su rendimiento.

En lo respecta al tercer objetivo de investigación, se han implementado tres herramientas informáticas, a saber, *PreProccesor*, *STSModule* y *DSMModule*. Así, *PreProcessor* es un programa que ayuda los usuarios a anotar datos de textos sin procesar. Aunque programas como *Part of Speech taggers*, *Lemmatisers*, *Stemmers*, *Named Entities Recognisers*, *Sentence Splitters*, *Tokenisers* o *Stopword Checkers* se pueden utilizar con este fin, se trata de aplicaciones independientes creadas para un propósito específico (por ejemplo, identificar la raíz de la palabra). Por lo tanto, cuando los usuarios desean usar más de uno o importarlos en sus propios programas o aplicaciones su integración debe ser completa y requiere de mucho tiempo. Para dar solución a este escollo, surge *PreProcessor*, que ofrece una variedad robusta

[22] `https://github.com/hpcosta/SCleaner`
[23] `http://www.lexytrad.es/scleaner/index.php`

y dinámica de funcionalidades morfosintácticas para anotar datos sin procesar aprovechando las bibliotecas de código abierto más conocidas del mercado. En segundo lugar, *STSModule* tiene como objetivo ayudar a los usuarios, mediante la combinación de varios recursos semánticos con métodos estadísticos, a calcular la similitud semántica entre oraciones o documentos en inglés, de gran importancia para una amplia variedad de aplicaciones de PLN o traducción automática. En tercer lugar, *DSMModule* surge con el propósito de ofrecer un programa de fácil manejo, capaz de medir y clasificar oraciones o documentos por su grado de similitud. En definitiva, permite determinar a los usuarios si un documento específico debe o no incluirse en el corpus en cuestión.

Asimismo, la presente tesis doctoral está avalada por trece contribuciones, previamente revisadas por pares, y publicadas en repertorios tanto nacionales como internacionales. Tres de ellas han reflejado interesantes hallazgos en torno al concepto de corpus comparable (Costa et al., 2014c; 2015d and Costa et al., 2015e). Concretamente, se han analizado las herramientas y tecnologías de compilación de corpus comparables más conocidas en el mercado, lo que ha permitido la identificación de sus principales limitaciones (Costa et al., 2014c; 2015d and Costa et al., 2015e). Así, a pesar de los esfuerzos por mantener estas herramientas actualizadas, se ha llegado a la conclusión de que, en la mayoría de los casos, esto no ocurre y la tecnología en la que se basan puede considerarse obsoleta. De esta forma, en un intento de demostrar la imbricación entre diversas disciplinas, a saber, ingeniería de software, experiencia de usuario y la lingüística computacional, se ha llevado a cabo la implementación de la herramienta de compilación de corpus comparables iCompile Corpora. El segundo objetivo principal de este trabajo se ha centrado, de manera general, en la evaluación de las tecnologías actuales utilizadas por traductores e intérpretes para la consecuente propuesta de alternativas de mejora. El resultado de estas investigaciones ha sido reflejado en cinco publicaciones, a saber, Costa et al., 2014a, Costa et al., 2014b, Costa et al., 2015b, Costa et al., 2016b and Costa et al., 2017. De ellas, cuatro están focalizadas tecnologías de la interpretación (Costa et al., 2014a;b; 2015b; 2017) y, la restante, se dedica a la investigación del grado de familiaridad de los traductores con las herramientas terminológicas (Costa et al., 2016b), con el propósito de facilitar a estos profesionales la elección de las herramientas más adecuadas a las necesidades específicas de un determinado encargo. Para ello, en primer lugar, nos hemos encargado de identificar las necesidades profesionales de los usuarios a través de resultados obtenidos en varias encuestas realizadas. A continuación, se ha estudiado el conjunto de funcionalidades ofrecidas por las herramientas de gestión terminológica más populares en el mercado para, seguidamente, plantear una nueva metodología estandarizada capaz de adaptar las características de estas herramientas a los requerimientos de traductores e intérpretes. En la misma línea, se ha seguido una metodología semejante para evaluar las herramientas de extracción terminológica en Costa et al., 2016b.

La tercera contribución principal de esta tesis está dedicada a determinar automáticamente la calidad de los textos que componen los corpus comparables (Zampieri et al., 2015; Costa et al., 2015a;c; Costa, 2015; Costa et al., 2016a), una de las cuales presenta varios enfoques para la discriminación entre idiomas y sus correspondientes variedades (Zampieri et al., 2015). La segunda contribución, a saber, Costa et al., 2015a, compara varios planteamientos para evaluar la similitud

entre las oraciones en español e inglés. Así, aunque el sistema no funcionó como esperábamos para el idioma español, ya que lo ubicó en el noveno puesto (de un total de 17), funcionó razonablemente bien para el inglés, ya que se ubicó en el puesto 33 (de un total de 74). Las tres publicaciones restantes (Costa et al., 2015c; Costa, 2015; Costa et al., 2016a) proponen varias metodologías capaces, no solo de describir automáticamente un corpus comparable, sino también de medir y comparar los diferentes conjuntos de documentos, así como agruparlos, a partir de su grado de parentesco de manera automática gracias al uso de técnicas de medias de similitud distribucional y PLN. En estos artículos, se ilustran los experimentos en cuestión realizados para demostrar que la metodología propuesta puede ser utilizada no sólo para clasificar documentos, sino también para describir y extraer información sobre corpus comparables y el grado de comparabilidad de sus documentos. Además, también evaluamos el rendimiento de las medidas de similitud distributiva en la tarea de filtrar el ruido documental; en este caso, documentos estaban fuera del dominio seleccionado. Aunque el coeficiente de correlación de rango de Spearman resultó ser incapaz de filtrar los documentos que conforman el ruido documental, desempeñó un papel importante en la descripción de los datos en cuestión; los coeficientes *Number of Common Tokens* (NCT) y el *Chi-Square* (χ^2), por su parte, demostraron ser eficientes en ambas tareas.

Aunque todas las publicaciones mencionadas aparecen citadas como referencia a lo largo de la tesis doctoral, y se puede acceder a su contenido completo en el Apéndice A (cfr. Appendix A), a continuación se exponen agrupadas de acuerdo a los objetivos de objetivos de investigación planteados (RQ) (cfr. Apartado 1.2).

- **RQ1**
  - **Costa et al. (2014c)**: Costa, H., Corpas Pastor, G., and Seghiri, M. (2014). iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, London, UK.
  - **Costa et al. (2015d)**: Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015). iCorpora: Compiling, Managing and Exploring Multilingual Data. In *7th Int. Conf. of the Iberian Association of Translation and Interpreting Studies, AIETI*, pages 74-76, Malaga, Spain.
  - **Costa et al. (2015e)**: Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015). Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora. In *New Horizons in Translation and Interpreting Studies (Full papers)*, pages 133-141, Geneva, Switzerland. Tradulex.
- **RQ2**
  - **Costa et al. (2014b)**: Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014). Technology-assisted Interpreting. *MultiLingual* #143, 25(3):27-32.
  - **Costa et al. (2014a)**: Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014). A comparative User Evaluation of Terminology Management Tools for Interpreters. In *25th Int. Conf. on Computational Linguistics*

(COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14), pages 68-76, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

- **Costa et al. (2015b)**: Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2015). An Interpreters' Guide to Selecting Terminology Management Tools. In *NATO Conf. on Terminology Management*, Brussels, Belgium.
- **Costa et al. (2016b)**: Costa, H., Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2016). Nine terminology extraction Tools: Are they useful for translators? *MultiLingual* #159, 27(3).
- **Costa et al. (2017)**: Costa, H. and Corpas Pastor, G. and Durán Muñoz, I. (2017) Assessing Terminology Management Systems for Interpreters. In Corpas Pastor, Gloria and Durán Muñoz, Isabel, *Trends in E-Tools and Resources for Translators and Interpreters*, volume 45, pages 57-84, Brill.

- **RQ3**
  - **Zampieri et al. (2015)**: Zampieri, M., Gebrekidan Gebre, B., Costa, H., and van Genabith, J. (2015). Comparing Approaches to the Identification of Similar Languages. In *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial'15). 2nd Discriminating between Similar Languages Shared Task (DSL'15)*, Hissar, Bulgaria.
  - **Costa et al. (2015a)**: Costa, H., Béchara, H., Taslimipoor, S., Gupta, R., Orasan, C., Corpas Pastor, G., and Mitkov, R. (2015). MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. In *9th Int. Workshop on Semantic Evaluation, SemEval'15*, pages 96-101, Denver, Colorado. ACL.
  - **Costa et al. (2015c)**: Costa, H., Corpas Pastor, G., and Mitkov, R. (2015). Measuring the Relatedness between Documents in Comparable Corpora. In *11th Int. Conf. on Terminology and Artificial Intelligence, TIA'15*, pages 29-37, Granada, Spain.
  - **Costa (2015)**: Costa, H. (2015). Assessing Comparable Corpora through Distributional Similarity Measures. In *EXPERT Scientific and Technological Workshop*, pages 23-32, Malaga, Spain.
  - **Costa et al. (2016a)** Costa, H., Durán Muñoz, I., Corpas Pastor, G., and Mitkov, R. (2016b). Compilação de Corpos Comparáveis Especializados: Devemos sempre confiar nas Ferramentas de Compilação Semi-automáticas? *Linguamática*, 8(1):17.

Esta investigación, asimismo, se ha desarrollado en el marco del proyecto europeo *EXPERT*[24] (*EXPloiting Empirical AppRoaches to Translation*), financiado por el Séptimo Programa Marco de Investigación y Desarrollo Tecnológico de la Unión Europea bajo el acuerdo de subvención número 317471. El objetivo del proyecto *EXPERT* es mejorar las tecnologías de la traducción existentes mediante el análisis

[24] `http://expert-itn.eu`

tanto de sus principales deficiencias como de las necesidades profesionales de los traductores e intérpretes. Habida cuenta de la relevancia tanto de la traducción humana como de la traducción automática en Europa, este proyecto apunta a unirlas a través del desarrollo de tecnologías de última generación que atiendan de los requerimientos profesionales de los traductores y de la política lingüística de la Comunidad Europea.

En este sentido, hoy en día, el progreso de la traducción automática es una necesidad innegable en el marco de actual entorno globalizado donde la comunicación multilingüe se vuelve cada vez más relevante. Así, los recientes avances en los sistemas de memorias de traducción, y de traducción automática han demostrado el potencial del enfoque para la producción de traducciones rápidas y de bajo coste. En definitiva, esta tendencia profesional en el ámbito de la traducción e interpretación se está convirtiendo en un recurso indispensable para el respaldo de la traducción humana. Por consiguiente, el proyecto *EXPERT* tiene como objetivo desarrollar nuevas tecnologías que aumenten la productividad y reduzcan los costes en el sector de la traducción y producción de contenido multilingüe (Orăsan et al., 2015).

Por su parte, el candidato al doctorado ha trabajado como investigador (*Early Stage Researcher*, ESR) en el proyecto mencionado anteriormente y fue responsable de investigar cómo se podían construir automáticamente repositorios de datos para asegurar su utilidad en múltiples enfoques de traducción e interpretación basados en corpus, así como para identificar problemas en las actuales herramientas asistidas por la tecnología y sugerir posibles mejoras entre ellas. En concreto, se ocupó de:

a) explotar las técnicas existentes para construir corpus comparables e investigar su utilidad para los sistemas de traducción automática y los usuarios de idiomas;

b) desarrollar técnicas para "limpiar el ruido documental" de los corpus comparables con el fin de convertirlos en una fuente de datos más fiable y útil tanto para los sistemas de traducción automática como para lingüistas;

c) utilizar técnicas de aprendizaje de transferencia para aplicar los conocimientos adquiridos en lenguas ricas en recursos con el fin de construir corpus para las lenguas pobres en recursos y los dominios específicos;

d) sugerir, o incluso crear, nuevas metodologías o herramientas para automatizar los procesos, aumentar la productividad y facilitar las actividades que requieren más esfuerzopara los lingüistas.

La nómina de integrantes de EXPERT, compuesta por seis universidades y varios socios comerciales, facilitó un sistema único para la formación, la colaboración y el intercambio de conocimiento entre los investigadores. Así, dentro de las actividades acometidas, se han realizado estancias en las diferentes instituciones que formaban parte del consorcio. Esta iniciativa ha permitido la realización de diversas actividades de investigación en dos instituciones internacionales participantes, lo que ha hecho posible la familiarización con otros enfoques y métodos teóricos desarrollados en el seno de los centros científicos en cuestión. En concreto, estas estancias han tenido lugar en la Universidad de Wolverhampton y en Translated. Así, la primera estancia internacional tuvo lugar entre septiembre de 2014 y diciembre de 2014 en la Universidad de Wolverhampton (Reino Unido), y más concretamente en el Instituto de Investigación en Procesamiento de la Información y

el Lenguaje (Research Institute in Information and Language Processing, RIILP[25]), uno de los grupos líderes en lingüística computacional en el Europa y que es reconocido, entre otras líneas de trabajo, por su especialización en la gestión y explotación de corpus. La estancia en Translated, empresa de traducción radicada en Roma (Italia) tuvo lugar entre octubre y diciembre del año 2015. En ella nos familiarizamos con la parte más profesional del mercado de la traducción y de la interpretación, entrando de lleno en el uso y evaluación de herramientas. Además de estas actividades de investigación, el candidato ha participado y estado involucrado en varios eventos y conferencias de formación, tales como conferencias y seminarios promovidos por el Programa de Doctorado en Lingüística, Literatura y Traducción de la Universidad de Málaga (UMA); sendos eventos de formación organizados por el proyecto de investigación *EXPERT*; diversas conferencias y ponencias internacionales sobre tecnologías de traducción, lingüística de corpus y PLN; así como en cursos especializados en la materia, como el de *Lisbon Machine Learning School* (LxMLS), entre otros ejemplos. Apasionado por la investigación, el candidato cuenta con más de ocho años de experiencia profesional en investigación, ha trabajado en diversos proyectos de vanguardia, tiene más de 30 publicaciones en distintos campos de conocimiento y ha colaborado en más de 20 eventos y encuentros científicos como miembro del comité de programa, revisor o editor.

Palabras Clave: corpus comparables, herramientas de extracción terminológica, herramientas de gestión de terminología, interpretación, lingüística computacional, lingüística de corpus, medidas de similitud distributiva, procesamiento de lenguaje natural, traducción automática, traducción humana.

[25] `http://rgcl.wlv.ac.uk/`

Chapter 1

Introduction

"I don't know anything,
but I do know that everything is interesting
if you go into it deeply enough."

—Richard Feynman

Textual corpora have long been the preferred resource in the language engineering and the linguistics communities. In language engineering, on the one hand, is mainly motivated by the need to use corpora as training data for Natural Language Processing (NLP, Jurafsky and Martin (2009)) applications, such as Machine Translation (MT, Poibeau (2017)) and Cross-Language Information Retrieval (CLIR, Meunier and Dymetman (2014)). In linguistics, on the other hand, corpora are of interest in themselves by making possible inter-linguistic comparisons and discoveries (Aston, 2016; Gil-Berrozpe and Faber, 2016). Indeed, it is generally accepted across both communities that corpora are a reliable alternative to lexicographical resources and dictionaries which may offer only limited coverage. Their objectivity, reusability, multiplicity and applicability of uses, easy handling and quick access to large volume of data are just an example of their advantages over other types of limited resources like thesauri or dictionaries. For instance, in the case of terminology new terms are coined on a daily basis and dictionaries or other lexical resources cannot keep up with the rate of emergence of new terms (Mitkov, 2016). Thus, terminologists seek to analyse the use and/or identify the translation of a specific term using corpora (Temmerman, 2000; Bouamor et al., 2013; Hazem and Morin, 2013; Faber, 2015). Moreover, the applicability of current data-driven methods directly depends on the availability of large quantities of parallel and comparable data.

Ideally, parallel data (i.e. collection of texts, each of which is translated into one or more other languages than the original (EAGLES, 1996b)) would be the best resource for both language engineering, such as NLP applications, and for language users, such as translators, interpreters and language learners (Cencini, 2002; Kotani and Yoshimi, 2015; Laviosa, 2016). Nevertheless, the lack of sufficient and up-to-date parallel corpora, Translation Memories (TM, Somers (2003)) or other parallel resources, especially for narrow domains and poorly-resourced languages is currently one of the major obstacles to further advancement across various areas like translation, language learning, automatic and assisted translation, amongst others (Skadiņa et al., 2010b). An alternative and promising approach would be to benefit from non-parallel bilingual and multilingual text resources, also known as comparable corpora (i.e. corpora that include similar types of original texts in one or more languages using the same design criteria (cf. EAGLES, 1996b; Corpas Pastor, 2001:158; Maia, 2003), which are easier and faster to compile.

Comparable corpora, whether "strongly comparable" by definition or "weakly comparable" (Skadiņa et al., 2010a) have been widely used as a resource in MT (Rapp et al., 2016), by professional translators (Corpas Pastor and Seghiri, 2009; Seghiri, 2015; 2017b; Arce Romeral and Seghiri, 2018b) or even by interpreters for interpreting research and learning (Cencini, 2002; Straniero S., 2012; Fantinuoli and Zanettin, 2015; Defrancq, 2016; Corpas Pastor and Seghiri, 2016; Seghiri, 2017a; Arce Romeral and Seghiri, 2018a; Pérez-Pérez, 2018). Indeed, comparable corpora can facilitate almost any multilingual application and can be beneficial to almost any language user. Thus, comparable corpora can be seen as the most versatile, valuable and practical resource for bi- and multilingual applications and research studies. Various examples of their applicability can be found, for instance at the workshop series on "Building and Using Comparable Corpora" (BUCC[26]), which has been promoting and exchanging progress in this exciting emerging field by bundling

[26] https://comparable.limsi.fr/bucc2016/

its research since 2007.

1.1 Research Problems

Even though comparable corpora can compensate for the shortage of linguistic resources and ultimately improve automated translations quality for under-resourced languages and narrow domains for example (Munteanu and Marcu, 2005; Eisele and Xu, 2010; Skadiņa et al., 2010a), the problem of data collection presupposes a significant technical challenge (Maia, 2003). Although the compilation process could be manually performed, nowadays specialised tools can be used to automate this tedious task (Baroni and Bernardini, 2004; Baroni et al., 2006; de Groc, 2011). Nevertheless, these compilation tools are scarce or proprietary, simplistic with limited features and designed to compile one monolingual corpus at a time, in other words they do not completely fulfil the users' needs (Costa et al., 2014c). Consequently, their simplicity, lack of features, performance issues and usability problems (Gutiérrez Florido et al., 2013) result in a pressing need of improvement or event to design new compilation tools tailored to fulfil not only translators' and interpreters' needs, but also professionals' and ordinary people's (Costa et al., 2014c; 2015d;e).

Comparable corpora is often used by translators and interpreters on their daily tasks, which by a way of example include terminology extraction and management. Nevertheless, interpreters have not benefited from the same level of automation or innovation as like translators, for whom a myriad of computer-assisted tools are available. (Durán Muñoz, 2012; Costa et al., 2014b; 2016b; Corpas Pastor and Durán Muñoz, 2017; Corpas Pastor, 2018). Their work relies by and large on traditional or manual methods. The solutions tailored to the interpreters' needs are few and still far behind, specially what concerns terminology tools (Costa et al., 2014a). Thus, there is also an urgent need to improve the current technology stack or even develop new methods to automate the process, increase the productivity and ease the labour-intensive activities of an interpreter before and during an interpreting service.

Another pressing issue is related with the uncertainty about the form and content of the comparable documents either manually or automatically compiled. Both form and content are of paramount importance in the construction of comparable corpora and in the optimal results during the analysis. The EAGLES - Expert Advisory Group on Language Engineering Standards Guidelines (EAGLES, 1996b) defined "comparable corpora" as follows: "A comparable corpus is one which selects similar texts in more than one language or variety. There is as yet no agreement on the nature of the similarity, because there are very few examples of comparable corpora.". Since 1996, when this definition was given, many comparable corpora have been compiled, analysed and employed in a wide range of disciplines, since it has become an essential resource in several research domains and Natural Language Processing (NLP) applications. Therefore, at this point, we can state that there are no more "very few examples of comparable corpora". Nevertheless, "there is as yet no agreement on the nature of the similarity" so far (Braschler and Scäuble, 1998; Maia, 2003; Bekavac et al., 2004; Fung and Cheung, 2004; Skadiņa et al., 2010a). The uncertainty about the data we are dealing with is still an

inherent problem to those who deal with this resource. Indeed, little work has focus on automatically characterising such linguistic resources (Kilgarriff, 2001; Sharoff, 2013; Köhler, 2013), and attempting a meaningful description of their content is often a perilous task (Corpas Pastor and Seghiri, 2009). Although some work has been carried out on assessing comparability of comparable corpora (Skadiņa et al., 2010b:162), there is yet no methodology to automatic describe, measure and rank documents according to their internal degree of relatedness.

1.2 Objectives and Approach

As stated in Maia (2003), "comparable corpora are seen as answering perceived needs for texts as examples of 'natural' original text in the source language culture" and, thus, we have witnessed an increased interest for these resources and a great boost of comparable corpora compilation in research in the last decades. Nevertheless, the current comparable corpora compilation tools still lack of some features (Gutiérrez Florido et al., 2013). By a way of example, they do not allow to compile multilingual comparable corpora at a time and they do not allow the use of more than one Boolean operator when creating search query strings (Baroni and Bernardini, 2004; Baroni et al., 2006; de Groc, 2011). In fact, not only in the compilation phase, but also in the exploitation phase, the current tools still do not fulfil all the users' requirements. Recent users' surveys (cf. Rodríguez and Schnell, 2009; Bilgen, 2011; Durán Muñoz, 2012; Zaretskaya et al., 2015) pointed out the need to investigate in more detail translators' and interpreters' attitudes towards terminology tools, especially the reasons that prevent the vast majority of professional to adopt them. Thus, it is of extreme importance to analyse, suggest improvements or even come up with new tools and methodologies capable of helping language users, such as translators and interpreters automatise their tasks (e.g. assembling, extracting and managing data, either mono- or multilingual).

Against the background of the increasing importance of multilingual tools, the first objective of this thesis is two-fold. First, analyse the best known compilation tools on the market, identify their limitations and propose new ways of improvement. Then, based on the pre-identified drawbacks and strengths, design and develop a novel, flexible and robust web-based prototype capable of exploiting both mono- and multilingual comparable corpora from the Web. In other words, create a prototype focused on increasing the flexibility and robustness of the compilation process by solving some of the usability problems found in the current compilation tools available on the market and by reducing their limitations and performance issues. Moreover, it is a priority to build a simple interface with easy-to-follow steps to allow not only experienced users build comparable corpora, but also laypersons with less experience in the area. It is important to mention that the main goal here is not to build a commercial tool, but instead build a proof-of-concept and suggest new directions to advance forward the comparable corpora compilation process.

The second main objective of this work is to address translators' and interpreters' needs and suggest new methodologies or tools to help them increase the productivity and ease their labour-intensive activities (mostly in the preparation stage of a given task). To do so, firstly, it is necessary to identify the users' requirements regarding the use of terminology tools, which can be done by analysing various users' surveys

in the literature. Then, examine the most known Terminology Management Tools (TMT) and Terminology Extraction Tools (TET) on the market with the purpose of identify the set of features these tools have to offer. Finally, by comparing the set of software functionality functions offered by these tools with the users' requirements, a new standardised methodologies capable of evaluating these tools can be proposed, as well as point out ways of improvement.

The notion of comparability is a loose one, and comparable corpora range from lowly comparable ones to highly comparable ones. Not only for data-driven Natural Language Processing (NLP) tasks but probably for all tasks relying on this type of resource, using better corpora often leads to better results. Although this point has largely been ignored in previous works on the subject, the third main goal of this work is to explore new methods capable of helping users accessing comparable corpora and filtering out irrelevant documents in an automatic way, and therefore improve the corpus quality. Decisions at the outset of compiling a comparable corpus are of crucial importance for how the corpus is to be built and analysed later on. Several variables and external criteria are usually followed when building a corpus but little is been said about textual distributional similarity in this context and the quality that it brings to research. In fact, one of the most important drawbacks when dealing with automatic compilation is noise (Costa et al., 2015c), that is, the amount of irrelevant information that is included in a corpus during the compilation. Prompted by this noise, researchers are forced to perform strict supervision to reduce it and, thus, avoid possible pitfalls during the subsequent analysis. It almost goes without saying that this process requires human intervention afterwards, which results extremely demanding when trying to get rid of these noisy-documents retrieved by the compiler. In this vein, this work aims at building simple, yet efficient methodologies capable of measuring and ranking documents based on their similarity scores. To do so, Distributional Similarity Measures (DSMs) will be combined with well-known NLP techniques in order to assess the corpus internal degree of relatedness. In the last instance, the resulting methodology will allow not only to measure and rank documents, but also to describe and extract information about the corpus in hand and the degree of relatedness in it.

To sum up, this thesis aims at addressing the three aforementioned main objectives, which have been summarised in the following Research Questions (RQ) and sub-questions:

RQ1: Is it possible to create a simpler and user-friendly comparable corpora compilation tool?

a) Is it feasible to allow the user to compile multilingual comparable corpora instead of one monolingual corpus at a time?
b) Can we solve some usability problems of the current compilation tools?
c) How to simplify the compilation process in order to fulfil not only translators' and interpreters' needs, but also the needs of other professionals and laypersons?

RQ2: How to identify the most suitable TMT and TET for a given translation or interpreting task?

a) Is it possible to identify the most required TMT and TET features based on users' surveys?

b) How to convert the most desirable TMT and TET features into a standardised scoring system?

c) What can be the possible improvements to be made in the current TMT and TET to fulfil the current translators' and interpreters' needs?

RQ3: How to automatically assess and measure the internal degree of relatedness in comparable corpora?

a) Is it possible to automatically assess the internal degree of relatedness between sentences, documents or even between corpora?

b) How can NLP and statistical methods be combined to build automatic methods capable of assessing and ranking sentences and documents according to the content they share between each other?

c) Can comparable corpora's internal quality be improved by filtering out documents with a low degree of relatedness?

1.3 Original Contributions

This thesis is composed by 13 previously published and peer-reviewed publications in national and international events. Three of them reporting original contributions in the comparable corpora research domain (see Costa et al., 2014c; 2015d and Costa et al., 2015e). In detail, these publications aim at analysing the best known compilation tools and methodologies used in both fields research and industry. After a careful analysis of the most known comparable compilation tools on the market, several limitations and drawbacks were identified and reported in Costa et al., 2014c; 2015d and Costa et al., 2015e. Despite of the extraordinary effort and time invested on these tools, we conclude that they are not keeping up to the current user's requirements, and the technology they are build on can be sometimes considered obsolete. In an attempt to show the research community that it is possible to fuse various disciplines, such as Software Engineering, User Experience (UX) and Computational Linguistics to create a compilation tool that tackles the current usability problems and performance issues found in the current tools on the market, a new web-based application prototype named iCompileCorpora has created. iCompileCorpora can be considered a reliable and intuitive web application that allows the user to compile multilingual comparable corpora intuitively. Some advantages over other tools on the market are, the option to build multilingual comparable corpora at a time, to make full usage of various Boolean operators while creating the searchable queries, amongst other non-visible improvements than meets the eye, such as performance, document formatting and user experience.

The second set of contributions of this work focused on the exploitation of new methodologies to assess and evaluate the current technologies used by translators and interpreters in their daily work and propose new ways of improvement. As a result, a set of five publications was written to cover this niche, namely Costa et al., 2014a, Costa et al., 2014b, Costa et al., 2015b, Costa et al., 2016b and Costa et al., 2017. Four of them focus more on technology-assisted interpreting

(Costa et al., 2014a;b; 2015b; 2017) and the remaining one investigates translators' attitudes towards terminology tools (Costa et al., 2016b). In an attempt to help interpreters and also translators choosing the best tool that best caters for their specific needs, we first focused on compiling the users' requirements by using various users' surveys. In parallel, we identified a set of features offered by the most known Terminology Management Tools (TMT) on the market. Then, by matching the software functionalities offered by these tools with the users' requirements, a new standardised methodology capable of evaluating TMT on the market was proposed. A similar methodology was followed to assess Terminology Extract Tools (TET) in Costa et al., 2016b. Finally, we suggested new directions of improvement for both types of tools, mostly do to the current displacement between the users' needs and offered software functionalities.

The third main contribution of this thesis focused on automatically assessing the quality of documents in comparable corpora (Zampieri et al., 2015; Costa et al., 2015a;c; Costa, 2015; Costa et al., 2016a). One of these five publications presents various approaches to discriminate between similar languages and language varieties (Zampieri et al., 2015). This work was submitted to the Discriminating between Similar Languages (DSL) shared task. We got 2^{nd} (out of 9 teams) on one test set and 4^{th} (out of 7 teams) on the other. The second publication, i.e. Costa et al., 2015a, compares various approaches to compute the similarity between sentences in Spanish and English. Although the system did not perform as we expected for Spanish as it ranked 9^{th} (out of 17), it performed reasonably well for English, where it ranked 33^{th} (out of 74). The remaining three publications (Costa et al., 2015c; Costa, 2015; Costa et al., 2016a) propose various methodologies capable of, not only automatically describing a comparable corpus, but also measuring and comparing different sets of documents as well as ranking them by their degree of relatedness in an automatic fashion. By using Distributional Similarity Measures (DSMs) and Natural Language Processing (NLP) techniques, we build a methodology capable of measuring and ranking documents based on their similarity scores. In these articles various experiments were performed to demonstrate that the proposed methodology can be used not only to measure and rank documents, but also to describe and extract information about comparable corpora and the degree of relatedness of its documents. Moreover, we also evaluated how DSMs perform in the task of filtering out noisy documents, in this case out-of-domain documents randomly selected from a different corpus. Although Spearman's Rank Correlation Coefficient resulted incapable of filtering out noisy documents, it played an important role describing the data in hand. And, the Number of Common Tokens (NCT) and the Chi-Square (χ^2) demonstrated to be efficient in both tasks.

Although, all these original contributions are referred by reference throughout the thesis, their full content can always be accessed in Appendix A. Rather than presenting them in a chronological order of publication, hereafter they are grouped by Research Question (RQ) (see section 1.2).

◇ **RQ1**

- **Costa et al. (2014c)**: Costa, H., Corpas Pastor, G., and Seghiri, M. (2014). iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, London, UK.
- **Costa et al. (2015d)**: Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015). iCorpora: Compiling, Managing and Exploring Multilingual Data. In *7th Int. Conf. of the Iberian Association of Translation and Interpreting Studies, AIETI*, pages 74-76, Malaga, Spain.
- **Costa et al. (2015e)**: Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015). Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora. In *New Horizons in Translation and Interpreting Studies (Full papers)*, pages 133-141, Geneva, Switzerland. Tradulex.

◇ **RQ2**

- **Costa et al. (2014b)**: Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014). Technology-assisted Interpreting. *MultiLingual* #143, 25(3):27-32.
- **Costa et al. (2014a)**: Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014). A comparative User Evaluation of Terminology Management Tools for Interpreters. In *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, pages 68-76, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.
- **Costa et al. (2015b)**: Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2015). An Interpreters' Guide to Selecting Terminology Management Tools. In *NATO Conf. on Terminology Management*, Brussels, Belgium.
- **Costa et al. (2016b)**: Costa, H., Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2016). Nine terminology extraction Tools: Are they useful for translators? *MultiLingual* #159, 27(3).
- **Costa et al. (2017)**: Costa, H. and Corpas Pastor, G. and Durán Muñoz, I. (2017) Assessing Terminology Management Systems for Interpreters. In Corpas Pastor, Gloria and Durán Muñoz, Isabel, *Trends in E-Tools and Resources for Translators and Interpreters*, volume 45, pages 57-84, Brill.

◇ **RQ3**

- **Zampieri et al. (2015)**: Zampieri, M., Gebrekidan Gebre, B., Costa, H., and van Genabith, J. (2015). Comparing Approaches to the Identification of Similar Languages. In *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial'15). 2nd Discriminating between Similar Languages Shared Task (DSL'15)*, Hissar, Bulgaria.

- **Costa et al. (2015a)**: Costa, H., Béchara, H., Taslimipoor, S., Gupta, R., Orasan, C., Corpas Pastor, G., and Mitkov, R. (2015). MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. In *9th Int. Workshop on Semantic Evaluation, SemEval'15*, pages 96-101, Denver, Colorado. ACL.

- **Costa et al. (2015c)**: Costa, H., Corpas Pastor, G., and Mitkov, R. (2015). Measuring the Relatedness between Documents in Comparable Corpora. In *11th Int. Conf. on Terminology and Artificial Intelligence, TIA'15*, pages 29-37, Granada, Spain.

- **Costa (2015)**: Costa, H. (2015). Assessing Comparable Corpora through Distributional Similarity Measures. In *EXPERT Scientific and Technological Workshop*, pages 23-32, Malaga, Spain.

- **Costa et al. (2016a)** Costa, H., Durán Muñoz, I., Corpas Pastor, G., and Mitkov, R. (2016b). Compilação de Corpos Comparáveis Especializados: Devemos sempre confiar nas Ferramentas de Compilação Semi-automáticas? *Linguamática*, 8(1):17.

As we can see, each RQ is addressed by more than one peer-reviewed publication. Table 1 summarises in which way the publications address the corresponding RQ. In brief, table 1 shows that the first RQ is addressed in the publications Costa et al. (2014c), Costa et al. (2015d) and Costa et al. (2015e), which describe various tools and methodologies used to compile corpora and also suggest various ways to improve the compilation process. The RQ2 is addressed in the publications Costa et al. (2014b), Costa et al. (2014a), Costa et al. (2015b), Costa et al. (2016b) and Costa et al. (2017). Although all of them are of interest for both translators and interpreters, the first four are more focus on the interpreters' needs and the last one on translators'. Finally, RQ3 is addressed in the publications Zampieri et al. (2015), Costa et al. (2015a), Costa (2015), Costa et al. (2015c) and Costa et al. (2016a). The first one focus on discriminating similar languages and the other four on assessing the internal degree of relatedness in corpora.

Apart from the aforementioned peer-reviewed publications, various programs and tools were created during this work. Table 2 introduces them and shows the addressed RQ. Briefly, there are two web-based tools associated with RQ1: the SCleaner, a web application that helps users to format text copied from a PDF file; and, the iCompileCorpora, a web interface that guides the user through the creation of comparable corpora. Regarding the RQ3, three programs were created: the PreProcessor, a program that helps users to annotate raw textual data; the STSModule, a program that aims at helping users computing the semantic similarity between sentences and documents in English; and, finally, the DSMModule, a program that helps the user to assess and rank documents according to their internal degree of similarity. It is important to mention that all the developed software was made publicly available and, thus free for being used by anyone, both in a research or in a commercial setting. We believe this is the best way to contribute for the advancement of science in general and Computational Linguistics (CL) in particular.

Publications	RQ1	RQ2	RQ3	Description
Costa et al. (2014c)	✓			Analyses the current comparable corpora compilation solutions' weaknesses and strengths and proposes various ideas of improvement.
Costa et al. (2015d)	✓			Justifies the need for better comparable corpora tools and proposes various ideas to improve their flexibility and robustness.
Costa et al. (2015e)	✓			Reviews the best known methods and tools used to compile parallel and comparable corpora and proposes new ideas to address their pitfalls.
Costa et al. (2014b)		✓		Offers a tentative catalogue of current language technologies for interpreters, divided into terminology tools for interpreters, note-taking applications for consecutive interpreting, voice recording applications and training tools.
Costa et al. (2014a)		✓		Presents an overview of the most relevant features that standalone TMTs should have in order to help interpreters before and during an interpretation service.
Costa et al. (2015b)		✓		Presents a set of measurable features that can be used to guide interpreters choosing the most adequate TMT for a given interpretation project and, briefly describes three TETs that could be used during the preparation stage to identify relevant terms from text.
Costa et al. (2017)		✓		Reviews the most up-to-date standalone, web-based and mobile TMTs specifically designed for interpreters and establishes a set of 15 measurable features that can be used to compare and evaluate them.
Costa et al. (2016b)		✓		Investigates translators' attitudes towards terminology tools and identifies a set of desirable features that can be used to help translators choosing the most adequate TET for a given task.
Zampieri et al. (2015)			✓	Presents, evaluates and compares various approaches to discriminate between similar languages and language varieties.
Costa et al. (2015a)			✓	Describes, compares and evaluates various approaches to compute the similarity between sentences in English and Spanish.
Costa (2015)			✓	Investigates and proposes a new methodology to automatically describe specialised comparable corpora.
Costa et al. (2015c)			✓	Proposes a simple methodology and studied various DSMs for the purpose of measuring the relatedness between documents and ranking them according to their degree of relatedness.
Costa et al. (2016a)			✓	Presents a detailed review of different statistical methods and NLP techniques to assess the internal degree of similarity in comparable corpora.

Table 1: Brief summary and Research Question addressed in each publication.

Name	RQ	Description
SCleaner [a,b]	RQ1	A web-based program that helps users to format text copied from a PDF file. When copying and pasting from a PDF file, users can find various formatting problems: white spaces, tabulations, sentence boundaries, etc. Scleaner removes extra tabs and white spaces, and splits sentences in the right place automatically.
iCompileCorpora [c,d]	RQ1	A web interface that guides the user through the creation of a web corpus. Designed for both novice and experts in the field, iCompileCorpora not only provides a simple interface with simplified steps, but also permits advanced users to set advanced compilation options during the compilation process.
PreProcessor [e]	RQ3	A program that helps users to process and annotate raw textual data. Despite various Part of Speech taggers, Lemmatisers, Stemmers, Named Entities Recognisers, Sentence Splitters, Tokenisers and Stopword Checkers can be used for this purpose, they are independent programs built for a specific purpose (e.g. identify the word's stem). Thus, when users want to use more than one or import them in their own programs/applications, their integration turns to be really complex and time-consuming. As an attempt to fulfil this gap, PreProcessor aims at offering the user with a simple, yet robust and agile variety of morphosyntatic options to process and annotate raw textual data by taking advantage of the best known open-source libraries on the market.
STSModule [f]	RQ3	STSModule (Semantic Textual Similarity Module) aims at helping users computing the semantic similarity between either sentences or documents in English. Similarity measures play an important role in a wide variety of NLP applications. By a way of example, IR relies on semantic similarity in order to determine the best result for a related query. Semantic similarity also plays a crucial role in other applications such as Paraphrasing and TM. However, computing semantic similarity between sentences and documents remains a complex and difficult task. As an attempt to fulfil this gap, STSModule aims at offering the user with a simple, yet very efficient approach to compute semantic similarity by combining various semantic resources with statistical methods.
DSMModule [g]	RQ3	DSMModule (Distributional Similarity Measures Module) aims at offering the user with a simple, yet efficient program capable of measuring and ranking either sentences or documents based on their similarity scores. Decisions at the outset of compiling a comparable corpus are of crucial importance for how the corpus is to be built and analysed later on. The DSMModule brings together methods from different areas of knowledge with the purpose of accessing, measuring and ranking documents based on their shared content, and consequently help researchers decide whether a specific document should be integrated in the corpus or not.

[a] `http://www.lexytrad.es/scleaner/index.php`
[b] `https://github.com/hpcosta/SCleaner`
[c] `https://github.com/hpcosta/iCompileCorpora`
[d] `https://icompilecorpora.herokuapp.com/home`
[e] `https://github.com/hpcosta/PreProcessor`
[f] `https://github.com/hpcosta/STSModule`
[g] `https://github.com/hpcosta/DSMModule`

Table 2: Developed software, brief summary and Research Question addressed.

1.4 Research Contextualisation

This research makes part of the European project EXPERT[27] (EXPloiting Empirical appRoaches to Translation), funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n.317471. The goal of the project is to improve existing translation technologies by addressing their most well-known shortcomings along with better consideration of user requirements and feedback, in order to improve translation quality and user satisfaction. As human and automatic translation are an important part of the policy of multilingualism within Europe, this project aims at bringing the two together through the development of next generation technologies to address the needs of both translators and EC policy.

Nowadays, automatic translation is an undeniable need in a globalised world where communication using several languages becomes increasingly more relevant. Translation Memory (TM) and Machine Translation (MT) systems are the two most elaborate technologies to support human translation. Recent developments in the area of Example-based and Statistical Machine Translation (EBMT, Poibeau (2017) and SMT, Koehn (2010), respectively) have shown the potential of data-driven approaches for producing fast and low cost translations. A number of user studies have however established shortcomings in the state-of-the-art of these technologies, including poor quality translations for low resource languages, interfaces that do not take into account user requirements and the user feedback. In order to improve current data-driven MT technologies (e.g. TM, SMT and EBMT) is it necessary to exploit their individual strengths through their combination and by addressing some of the main limitations of each of these technologies. Thus, the EXPERT project aims to develop new technologies that can increase productivity and reduce costs in the translation sector, as well as facilitate reliable communication and content creation in multiple languages (cf. Orăsan et al. (2015)).

The PhD candidate worked as a Early Stage Researcher (ESR) in the aforementioned project and he was responsible for investigating how data repositories could be automatically built to ensure their usefulness to multiple corpus-based approaches to translation and interpreting and, for identifying possible pitfalls in the current technology-assisted tools and suggest improvements. Which included: i) exploiting existing techniques for building comparable corpora and investigating their utility for MT systems and language users; ii) developing techniques for "de-noising" comparable corpora to make it a more reliable and useful data source for both MT systems and language users; iii) using transfer learning techniques to apply the knowledge learned for resource-rich languages in order to build corpora for resource-poor languages and narrow domains; iv) suggesting or even creating new methodologies and/or tools to automate processes, increase productivity and ease the labour-intensive activities of language users, such as translators or interpreters.

Consisting of six universities and various commercial partners, the project offered a unique infrastructure for training, collaboration and exchange of experience between the researchers. Thus, research activities within this project included various visits (secondments) to other institutions within the EXPERT consortium.

[27] `http://expert-itn.eu`

Thanks to that, the candidate carried out various research activities in two international institutions. These visits aimed to give him the possibility to better understand and practice the subject of study in a commercial context, as well as get acquainted with various theoretical approaches and methods developed by other research institutions. In detail, the candidate visited two institutions, the University of Wolverhampton and Translated s.r.l. The first secondment took place on September, 2014 until December, 2014 in the University of Wolverhampton, more precisely in the Research Institute in Information and Language Processing (RIILP)[28], which is one of the top leading groups in Computational Linguistics (CL) in the UK and well-known by their particular specialism in corpus development and exploitation. There, the candidate had the opportunity to improve his communication and acquire complementary skills in core research areas, such as CL and NLP. RIILP was a perfect place to work on data collection and test his findings on automatic corpora compilation. Regarding his second secondment, in Translated s.r.l. (between October, 2015 and December, 2015 in Rome, Italy), the candidate had the opportunity to receive local training in an industrial environment. Being a leading language service provider and translation technologies developer, Translated s.r.l. provided an excellent environment to work on the infrastructure for data collection and evaluation.

Apart from these main research activities, the candidate participated and was involved in various training events and conferences, such as: local conferences and seminars within the PhD programme at the University of Malaga (UMA) –his host institution; various training events organised by the EXPERT consortium, which provided ESRs and Expert Researchers (ERs) with knowledge and the necessary skills to fully carry out their professional research; various international conferences in translation technologies, CL and NLP; as well as in specialised courses like the Lisbon Machine Learning School (LxMLS).

Being passionate for research, the candidate counts with more than 8 years of experience in research, worked in various cutting-edge projects, has more than 30 publications in various fields and helped in more than 20 scientific events as a programme committee member, reviewer and/or editor. His main research interests lie in CL and Artificial Intelligence (AI), especially their practical applications in the fields of Translation Technologies, Specialised Translation, NLP, Information Extraction (IE) and Information Retrieval (IR). Apart from that, he is also interested in (and have worked on) a number of other topics, such as Recommender Systems (RS), Multiagent Systems, Affective Computing, amongst others. The three years he worked as a ESR gave him the opportunity to meet, work and learn from extraordinary professionals that he encountered all over the world. Autodidact, result-oriented, self-driven, highly motivated, creative, with a strong analytic skills, always looking for simplicity, efficiency and ways of self-improvement, the candidate used his hungry for learning and advance science forward to create and suggest new ways to push forward the current translation and interpreting technologies. Moreover, this thesis somehow reflects how his computer science skills contributed to MT in general and CL in particular.

[28] http://rgcl.wlv.ac.uk/

1.5 Outline of the Thesis

After the introductory and background knowledge chapters, in which the research context, problems, approaches and resulted contributions are briefly described and, the theoretical contexts of this work are outlined, respectively (Chapters 1 and 2), each one of the next chapters focus on one of the three main Research Questions (RQs). Besides describing in detail the followed procedures, one or more experiments towards its validation are reported in each chapter (Chapters 3, 4 and 5). Finally, after the concluding remarks (Chapter 6), in the end of the thesis, one appendix was specifically created to include all the resulted scientific publications associated with this work (Appendix A). Hereafter, each chapter is briefly summarised.

Chapter 1 introduces the research context, problems, goals, approaches and summarises the resulting contributions.

Chapter 2 explains (mostly) theoretical background knowledge that supports this research. It starts by formalising the concept of corpus. Then, various compilation design and protocol stages are described in detail. Given that on of the goals is to propose new methods for assessing the internal degree of comparability in comparable corpora, the last section is dedicated to this topic.

Chapter 3 focus on different aspects of the first RQ. In detail, this chapter aims at presenting a user-friendly web-based application capable of retrieving comparable data from the Web. To do so, firstly, the shortcomings and strengths of the most well-known comparable corpora compilation tools available on the market are analysed. Then, with the aim of solving some of their performance, usability and design problems, an innovative multilingual web-based comparable corpora compilation prototype, named iCompileCorpora is presented. Finally, some ideas for further improvements are given in the end of the chapter.

Chapter 4 explores and proposes various methods to assess interpreters' and translators' terminology tools. With the purpose of exploring the second RQ, this chapter starts by offering a tentative catalogue of technology-assisted interpreting tools for interpreters. Then, the following section highlights some of the features that interpreters expect from a Terminology Management Tool (TMS) and proposes a standardised scoring system to evaluate current TMS available on the market. Next, the third part of this chapter focus on Terminology Extraction Tools (TET) for translators. After identifying the priorities for the design and features to be included in a TET, a comparative analysis of various well-known TET currently available on the market is made. Finally, the last section presents our main findings and highlights some ideas to improve the current interpreters' and translators' terminology tools.

Chapter 5 aims at exploring the various aspects of the third RQ. Namely, it describes a methodology for automatically assess, measure and rank documents accordingly to their internal degree of relatedness in comparable corpora. In detail, this chapter starts by presenting some theoretical background knowledge and related

work on the field. Then, an overview of our methodology is meticulously described, together with the various NLP techniques and statistical methods involved. Next, various experiments and corresponding results are reported and analysed in detail. The last section focus on discussing our general findings and on suggesting future research directions.

Chapter 6 presents a final discussion on this research and highlights its main contributions. In the end, some cues are given for further improvements and additional work.

Appendix A reproduces the resulted scientific publications of this research in a chronological order.

Chapter 2

Background Knowledge

"If I have seen further
it is by standing on the shoulders of giants"

—Isaac Newton

Corpus linguistics is the study of language, which uses a collection of "real world" texts called corpus to analyse and investigate various linguistic questions (McEnery et al., 2006; Taylor, 2008; Lüdeling and Kytö, 2008), such as how language varies from place to place, determine how specific words and their synonyms collocate and vary in practical use, amongst other linguistic tasks that will be addressed later on. As pointed out in Taylor, 2008:180, corpus linguistics can be seen as "a tool, a method, a methodology, a methodological approach, a discipline, a theory, a theoretical approach, a paradigm (theoretical or methodological), or a combination of these" –in other words, a truly versatile area of knowledge. Due to the fact that this area offers an unique view to the language dynamism, it is not surprising that corpus linguistics is one of the most widely used methodologies since the early 20^{th} century (Firth, 1935) and, one of the preferred resource in the language engineering and the linguistics communities. Language engineering is mainly motivated by the need to use corpora as training data for Natural Language Processing (NLP) applications, such as Machine Translation (MT) systems. In linguistics communities, on the other hand, corpora are of interest in themselves by making possible inter-linguistic comparisons and discoveries. Indeed, corpora can facilitate almost any multilingual application and can be beneficial to almost any language user. Thus, corpora can be seen as the most versatile, valuable and practical resource for monolingual, bilingual or even for multilingual applications and research studies. Although the term corpus has been used as a general term to define any compilation of textual data, a collection of texts can not be considered a corpus if both a set of clear design criteria is not established and a systematic compilation protocol carried out a priori.

This chapter introduces (mostly) theoretical background knowledge that supports this research. More precisely, it reproduces and explains in detail various concepts of corpus. Firstly, section 2.1 formally defines the concept of corpus. Then, the corpus design criteria are described in section 2.2. Section 2.3 presents the various compilation protocol stages. Given that our work aims at proposing new methods to assess the comparability in comparable corpora, the last section is dedicated to this topic. In the end, we add some remarks in order to connect the described background knowledge with the work developed in the scope of this thesis (section 2.5). We decided to keep this chapter more theoretical, while the next chapters describe practical work, including existing tools and methodologies as well as related works on this field.

2.1 Definition of Corpus

Even though the term *corpus* has been used as a general term to define any compilation of textual data, a collection of texts is not *per se* a corpus. To be considered a corpus in the strict sense of the term, a set of clear design criteria must be established and a systematic compilation protocol carried out (EAGLES, 1994; 1996b;c; Corpas Pastor, 2001), see sections 2.2 and 2.3 for more details. Although formalising the concept of corpus is not an easy task, the definition proposed by John Sinclair in EAGLES, 1996c:4 is the most accepted in the research community: "a corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language". In his work, Sinclair also defines the minimum criteria to be met by collections of texts, in electronic format, so these collections can be considered a proper corpus, namely: the quantity (the corpus size in number of words); the quality (representativeness and balance); the encoding simplicity; and, documentation (EAGLES, 1996c). Thus, a corpus should not be confused with other electronic collections, such as the *archive/collection* or the *electronic text library* (Atkins et al., 1992; Torruella and Llisterri, 1999:51-52).

- ◇ **Archive/Collection**: is a repository of readable electronic texts, not linked in any coordinated way, i.e. does not have any structure or linguistic criteria because the most important factor to its creation is the availability of the data.

- ◇ **Electronic text library**: is a collection of electronic texts in a standardised format with certain conventions related to the content, but without rigorous selectional constraints.

- ◇ **Corpus**: is a compilation of texts, but different to the electronic collections, a corpus attends to specific linguistic criteria. It is codified following a standardised and homogeneous process, allowing the study of the behaviour of one or more languages. Using Sinclair words, a "computer corpus is a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks" (EAGLES, 1996c:5).

It is also important to mention that a corpus can be divided into two more levels, *subcorpus* and *component* (cf. EAGLES, 1996b:5; Torruella and Llisterri, 1999:52).

- ◇ **Corpus**: as previously mentioned, a corpus can be seen as a set of pieces of language, selected and ordered according to explicit linguistic criteria with the purpose of representing the language or some part of it (EAGLES, 1996c:4).

- ◇ **Subcorpus**: a subcorpus inherits all the properties from the corpus as it is a part of a larger corpus.

- ◇ **Component**: a component is not necessarily an adequate sample of a language. Instead, it can be seen as a collection of pieces of the language that are selected and ordered according to a set of criteria with the purpose of characterise its linguistic homogeneity[29]. Whereas a corpus may illustrate

[29] Homogeneity: the quality of being similar or comparable in kind or nature.

heterogeneity[30], as well as a subcorpus to some extent, the component only illustrates a particularity of the language (EAGLES, 1996c).

Regarding the number of languages in a corpus, it can be called *monolingual* or *multilingual* corpus. In detail, a monolingual corpus is one that contains texts one language, while multilingual corpora contain texts in two or more languages. More precisely, a corpus composed by documents in two languages is called *bilingual* corpus, and when more than two languages are presented in the corpus it is called multilingual corpus.

2.2 Corpus Design and Classification

Mostly due to the direct dependency of current data-driven methods for large quantities of parallel and comparable data, now, more than ever, we are witnessing an increased interest for these types of resources. Nevertheless, their design and classification are a paramount importance in the construction of these resources and in the optimal results during their usage. Despite the absence of a well-defined design and classification criteria, one of the most complete proposals in the literature can be found in Corpas Pastor (2001) –see also Seghiri, 2006; Corpas Pastor, 2008; Corpas Pastor and Seghiri, 2009; Seghiri, 2011. In this work, the author combines various ideas proposed by several experts in the field (cf. EAGLES, 1994; 1996b; Baker, 1993; Johansson and Oksefjell, 1998; Torruella and Llisterri, 1999). In detail, in Corpas Pastor (2001), the author uses the EAGLES reports as a starting point (EAGLES, 1994; 1996b), extends the typology proposed in Torruella and Llisterri (1999), and merges it with the ideas proposed in Baker (1993) and Johansson and Oksefjell (1998) about multilingual corpus classification, in an attempt to establish a set of well-defined design and classification criterion. Hereafter, we reproduce and explain in detail the identified aspects of corpus design and classification, i.e. all the aspects related with their *size*, *specificity*, *sample size*, *encoding* and *documentation* (sections 2.2.1, 2.2.2, 2.2.3, 2.2.4 and 2.2.5, respectively).

2.2.1 Corpus Size

The first classification criterion is related with the percentage and distribution of the different types of text included in the corpus.

- ◇ **Large Corpus**: its size is not limited and it is usually composed by a large number of words. Another particularity of this type of corpus is their lack of representativeness and its unbalance sample sizes.
- ◇ **Balanced Corpus**: integrates several language varieties, in similar percentages.
- ◇ **Pyramidal Corpus**: the assembled texts are distributed by levels. These levels are characterised by the progressive increasing complexity of the topics included. In other words, the more complex the text is, the higher its level in the pyramid and more reduced the number of texts will be.

[30] Heterogeneity: the quality of being diverse and not comparable in kind.

- ◇ **Monitor Corpus**: the number of texts in this corpus is invariable, but constantly updated, i.e. old texts are replaced by new texts whenever possible. Thus, this corpus can be seen as a reference for the language evolution. Bowker and Pearson, 2002:12-13 named this corpus as *open corpus* due to its dynamism, and pointed out that "given the dynamic nature of Language for Special Purposes (LSP) and the importance of staying abreast of current developments in the subject field, open corpora are likely to be of more interest for LSP users".

- ◇ **Parallel Corpus**: is composed by collections of texts in one original language and its translations to one or more target languages. When only two languages are involved, i.e. when the corpus has the original texts and its translation to a single target language, it is named bilingual parallel corpus. When more than two target languages are involved it is named multilingual parallel corpus. The most well known multilingual parallel corpus, at least in Europe, is the Europarl Corpus (Koehn, 2005).

- ◇ **Comparable Corpus**: is a corpus that includes similar types of original texts. As it is compiled from a original language in accordance with the same design criteria, these texts allow the comparison of their interlingual components (Corpas Pastor, 2001:158). Similarly to the parallel corpus, when only two languages are involved the corpus is named bilingual comparable corpus and multilingual comparable corpus when more than two languages are involved. In addition to these two subtypes, a third one named monolingual comparable corpus was been proposed by Corpas Pastor, 2001:158. Different from the first two subtypes this specific corpus includes original texts and their translated texts in the same language.

2.2.2 Corpus Specificity

The second classification criterion classifies the corpus based on the included text specificity.

- ◇ **General Corpus**: as described by Bowker and Pearson, 2002:11-12, a general corpus is a corpus that "can be taken as representative of a given language as a whole and can therefore be used to make general observations about that particular language". As its main focus is the language for general purpose, i.e. the language used by ordinary people in everyday situations (Bowker and Pearson, 2002:12), a good example of a general corpus is a corpus composed by daily news or newspapers articles. Nevertheless, Corpas Pastor, 2001:156 clarifies that besides general corpus there are also restricted corpus, such as *specialised*, *generic*, *canonical*, *chronological* and *historical corpus*. The author also points out that a general corpus should not be confused with lower levels of corpus as the subcorpus or the component.

- ◇ **Specialised Corpus**: is a corpus that is focused on a particular aspect of a language (Bowker and Pearson, 2002). Using Bowker and Pearson, 2002:12 words, "it could be restricted to the Language for Special Purposes (LSP) of a particular subject field, to a specific text type, to a particular language

variety or to the language used by members of a certain demographic group (e.g. teenagers)".

- ◇ **Generic Corpus**: is a corpus that assembles samples from a particular gender.
- ◇ **Canonical Corpus**: is a corpus that contains complete works of an author.
- ◇ **Chronological Corpus**: is a corpus that contains texts that have occurred over a period of time. This type of corpus can be also referred as *synchronic corpus* (Bowker and Pearson, 2002:12).
- ◇ **Historic Corpus**: a corpus that includes texts from different periods of time with the purpose of carry out studies about the language evolution (Abaitua, 2002).

2.2.3 Corpus Samples Size

The third classification criterion is related with the quantity of text used in the samples.

- ◇ **Textual Corpus**: with the purpose of representing the language, as well as their most important varieties, a textual corpus is composed by *whole texts*, i.e. complete texts. This type of corpus is broadly used in the creation of grammars and dictionaries, for example.
- ◇ **Reference Corpus**: whereas textual corpus assembles whole texts, a reference corpus is composed by samples of the whole text, i.e. parts of it. The aim is not in the text itself, but rather it seeks to represent some particularity of a language or language characteristic.
- ◇ **Lexical Corpus**: built for a specific purpose, the lexical study, this corpus is composed by small samples with similar length.

2.2.4 Corpus Encoding

The fourth classification criterion is related to the corpus encoding.

- ◇ **Annotated Corpus**: in addition to the original texts, an annotated corpus also comprises information about some linguistic analysis, which can be for example tagsets for encoding linguistic annotation, such as segmentation of the text into sentences and words, morphosyntactic tagging, parallel text alignment, etc. By a way of example, the Corpus Encoding Standard (CES)[31] offers a set of encoding standards for corpus-based works. CES specifies the minimal encoding level that a corpus must achieve to be considered standardised.
- ◇ **Unannotated Corpus**: most often created for non-linguistic purposes, such as publishing. This raw text corpus presents a high level of simplicity since has not been added any type of linguistic annotation. The most common format is plain text.

[31] `http://www.cs.vassar.edu/CES`

2.2.5 Corpus Documentation

The fifth classification criterion is related with the corpus documentation.

- ◇ **Corpus with documentation**: to make best use of a corpus it is necessary, not only have access to the texts, but also to the explanatory documentation, such as the licence agreements and meta-data[32] information, also known as corpus manifest. As far as possible, all such supporting documentation should be included along with the corpus itself. Usually the structure of a document is divided into two elements, the header that contains the meta-data and, the body with the document content. For instance, the header can contain the following fields: title (the title of the document), author (the author of the document), year (publishing year), availability (free, license, etc.), amongst others elements that help to describe the document origin and structure. The document body contains text-entities and can also have sections. The basic text-entities can be lists, tables, paragraphs or other unformatted text. The sections have the purpose of separate the text-entities. There is a consortium named Text Encoding Initiative (TEI)[33] which purpose is the development and maintenance of a standard for the representation of texts in digital form.

- ◇ **Corpus without documentation**: as its name suggests, this type of corpus does not have any documentation associated.

2.3 Corpus Compilation Protocol

After establishing the design criteria, the next stage in the compilation process passes by defining the protocol. As proposed by Seghiri (2006) –see also Seghiri, 2008; Corpas Pastor, 2008; Corpas Pastor and Seghiri, 2009; Seghiri, 2011 and Seghiri, 2015– the compilation protocol can be divided into four steps: *finding data*, *downloading the data*, *normalisation* and *storage* (sections 2.3.1, 2.3.2, 2.3.3 and 2.3.4). An additional step should also be considered in order to ensure the *representativeness* of the samples, i.e. to determine whether the corpus is representative, or not, to the object of study (section 2.3.5).

2.3.1 Finding Data

The first stage consists in locating and accessing data available on the Internet. There are basically two types of searches that can be made over the Internet to find textual data, *institutional* and *thematic*.

The institutional search is directed to institutional companies, organisations and institutions. The information available through these specialised web sources result in a high standard of quality and reliability as the writers are professionals and specialists in the field.

The thematic search is normally carried out by the use of search engines. Firstly a set of keywords is defined. Then, these keywords are combined along with truncations and Boolean operators with the purpose of creating domain-specific

[32] Meta-data: data that describes other data.
[33] `http://www.tei-c.org/index.xml`

search queries. It is very important to create well-define queries in order to avoid retrieving large amount of irrelevant documents. Finally, the documents returned by these queries are manually or automatically analysed and the irrelevant ones are filtered out.

2.3.2 Downloading Data

Once defined the target sources, the next stage is to retrieve the data from these sources. As mentioned before, this process can be manual or automatically performed through specialised programs. Hereafter a short explanation about the main approaches used to acquire data is presented.

- ◇ **Existing Collections**: this approach takes advantage of existing collections, handcrafted or automatically created. If on the one hand these collections provide an instant availability of linguistic data, on the other hand they are limited to its design constrains, resulting in a static and obsolete resource for specific demands. The Portuguese newspaper CETEMPúblico (Santos and Rocha, 2001) and the Spanish CREA[34] are just two examples of corpora already collected and publicly available for consultation.

- ◇ **Web-based Approach**: this approach overcomes the problems in the previous approach, by taking advantage of all the resources available on the Internet. Despite its many advantages over existing collection, it has some drawbacks. Some of the advantages are the availability of massive amounts of electronic text, public domain documents, and wide reach of text-types/topics/genres/domains. The disadvantages are: the difficulty of copyright ascertainment (something that also occurs with the previous approach); additional effort to clean the documents' meta-data; the difficulty to achieve a balanced corpus; and, finally, despite of the quantity of information at our disposal, the difficulty in retrieving documents with high quality increases. Despite the drawbacks, this approach is widely used, not only by language users, but also by data-driven technology. Usually, one of the two web-based approaches is used: *Web Search Engine* or *Web Focused Crawling*.

  - ○ **Web Search Engine**: the aim of this approach is to search the Web for pages that contain information about a pre-defined topic (yet, it can be used to exploit corpus for broad topics or domains). To do so, a well-defined set of keywords that characterise a specific topic/domain should be defined. Then, these keywords are converted into search query strings. With the purpose of creating more accurate search queries, the keywords are combined with Boolean operators in order to define relationships between them. The next step is to submit these search query strings to a search engine. The quantity and quality of the retrieved documents completely depends on both the search queries and the search engine used (e.g. Google[35], Yahoo[36] and Bing[37]). Two

[34] http://www.rae.es
[35] https://www.google.com
[36] https://search.yahoo.com
[37] https://www.bing.com

examples of semi-automatic comparable corpora compilation tools that use this approach are BootCaT[38] (Baroni and Bernardini, 2004) and WebBooTCat[39] (Kilgarriff et al., 2004).

- **Web Focused Crawling**: this approach uses a specific type of program, named focused crawler. A focused crawler is a program created to retrieve data from the Web, but instead of submitting multiple queries to a specific search engine, a focused crawler selectively searches for web documents (pages) belonging to a specific topic by employing the hyperlink structure of the Web, i.e. the URL. In detail, the web crawling process starts with a set of pre-defined URLs. Usually, the crawler connects to a specific server or to a pre-defined set of URLs and starts the search process from it. Before starting the actual crawl process, domain-specific vocabularies are handcrafted or semi-automatically gathered from these web pages (for all the wanted languages). These vocabularies are very important in the process as they are used to find the seed URLs of the crawl, and consequently the "driver queries"[40] to steer the crawling process to pages that contain the wanted topic/domain. To settle a set of seed URLs for each language, the gathered vocabularies is queried in a search engine, e.g. Bing, Yahoo or Google, and the resulted URLs are used as seed URLs. Then, a priority queue that holds the URLs of the to-be-visited pages is initialised with these seed URLs. It is at this point that the actual crawl process starts. One by one, the head URL of the URL queue is removed and the page pointed by the URL is visited. The data inside the page is extracted and the language of the page is automatically detected. If the language is one of the wanted ones, the page content is matched against the driver query. If the math between the page and the driver query similarity exceeds a threshold, the page content is saved and, the out-links of each fetched page are extracted, scored and prioritised according to some pre-defined rules. Then, the crawling process continues until it comes to a dead end or until some restriction defined in the crawling policy is met. The set of policies could be the maximum number of pages to crawl, the page domain, the page language, amongst others. As a result, this approach is capable of locating large amounts of relevant documents on a particular topic/domain, as well as effective in automatically discarding irrelevant documents. By a way of example, a topical crawling approach can be used when corpora are needed to compensate for the limitations of general resources, such as general-purpose dictionaries, which do not cover vocabulary for special domains. As this approach is limited to a pre-defined topic/domain and vocabulary, it retrieves more accurate results, but compared to the previous approach requires an additional effort and more computational power. Two examples of parallel text mining systems that use this approach are BITS (Ma and Liberman, 1999) and PTMiner (Chen and Nie, 2000).

[38] `http://bootcat.sslmit.unibo.it`

[39] `http://sketchengine.co.uk`

[40] A driver query is a specific type of query containing the topic/domain vocabulary of a particular language.

2.3.3 Normalisation

The retrieved documents can be codified in a wide variety of file formats, such as HTML (*.html*), PDF (*.pdf*), Microsoft Word (*.doc*, *.docx*), etc. In order to make these documents usable by a corpus management tool, they need to be normalised to an acceptable format (the most widely used is plain text (*.txt*) with the character-encoding scheme ASCII or UTF-8). It is also important to take into account that some of these documents could contain information about one or more aspects of the data, such as descriptive or structural (e.g. HTML *tags*). In this case, they should be excluded from the original documents otherwise they will influence the results during the analysis. As Sinclair, 1991:21 pointed out, "the safest policy is to keep the text as it is, unprocessed and clean of any other codes".

2.3.4 Storage

The last compilation stage is the data storage. Despite the apparent triviality of this task, the correct storage allows to quickly access and retrieve the documents properly. The most common way of doing this is through the use of a root directory, where the files correctly identified are well-organised into folders and subfolders, also well-identified.

2.3.5 Representativeness

This additional stage should be considered in order to determine whether the samples are representative, or not, to the object of study (Lavid López, 2005). As mentioned by Biber, 1988:246, "the representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalisability of the results of the research". Furthermore, he also emphasises that "a corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language" (Biber, 1988:246). Although he remains conscious of the difficulties involved in compiling a corpus that could be defined as representative of a particular linguistic feature (Biber, 1988), the truth is that even today the concept of representativeness is still surprisingly imprecise, considering its acceptance as a central characteristic that distinguishes a corpus from any other kind of collection. Moreover, despite some authors agree with the importance of the quality and representativeness of the samples used to compile a corpus (cf. Biber, 1988; 1990; 1993; 1995; Atkins et al., 1992; Quirk, 1992; EAGLES, 1994; 1996b;c), still exists a surprising scarcity of studies devoted to analyse the quality and representativeness of a corpus. As pointed out by Flowerdale, 2004:18: "several corpus linguists have raised issues concerning the size and representativeness of specialised corpora as well as the generalizability of their findings. In fact, these are thorny issues which have also been widely debated in the literature on corpus studies in general, and to which there seem to be no easy answers".

In an attempt to solve this problem, Seghiri (2006) presents, for the first time, a method to quantify, a posteriori, the minimum number of documents and words that should be included in a specialised language corpus (see also Corpas Pastor and Seghiri, 2007a;b). Afterwards, it is not possible to establish the minimum number of documents for a given corpus a priori because the size will always depend on the language and text types involved (Corpas Pastor and Seghiri, 2007a:171). In their

work, the authors used the N-Cor algorithm to create an application named ReCor. Still without a consensus in the research community, ReCor can be seen as a good starting point for future research on this controversial task.

2.4 Comparability Degree in Comparable Corpora

Several variables and external criteria are usually followed when building a corpus but little has been said about textual distributional similarity in this context and the quality that it brings to research. The uncertainty about the content of the data still is an inherent problem to those who deal with comparable corpora in their research. Little work has been done on characterising such linguistic resource, and attempting a meaningful description of its content is often a perilous task. In theory, when building a comparable corpus, we expect to achieve a balanced corpus in terms of quantity and quality of its documents, but in practice, this phenomenon is pretty difficult not only to achieve, but also to measure. The criteria to define comparability is not universal and they always rely on the type of comparable corpus we want and the task we want to use the corpus for. Moreover, depending on the purpose a comparable corpus is build for, some features might be more important than others. The next two sections present various proposals and features used in the literature to compile, assess and classify comparable corpora according to their documents degree of similarity or comparable content (sections 2.4.1 and 2.4.2).

2.4.1 Features Selection

In order to assess the degree of comparability of a comparable corpus, a number of features need to be selected. The choice of these similarity features is influenced by different factors, such as the aim for which a comparable corpus is built for, or even the methodology employed for its acquisition. The criteria to define comparability are not universal and it always rely on the type of comparable corpus we want and the task we want to use the corpus for. Amongst the literature there are two main types of works on comparable corpora, which induce different choices (Goeuriot et al., 2009:56), *general language works* (where the documents usually share a domain and a period) and *specialised language works* (where choice of criteria is various). By way of example, a comparable corpus made of news articles will tend to focus on *publication dates*, in addition to the *domain* and *topic*. Given this example, it is easy to understand that some parameters certainly have precedence over others, always depending on the purpose that the corpus is built for. In the following paragraphs other features are presented, along with a short description about their relevance to the task they were used for.

Regarding on content, Morin et al. (2007) suggest that, for the task of terminology extraction the *quality* of a comparable corpus might be more important than its *size*. In their work, they reported better results with a smaller corpus if both subcorpora belong to the same *register* (Morin et al., 2007:671). Thus, the genre or register can be considered an important criterion to weight comparability. Goeuriot et al. (2009), apart from topic and domain, also considered the *type of discourse*, which proved to increase the degree of comparability between the documents.

Still in the context of terminology extraction, Leturia et al., 2009:55 consider the domain and topic similarity more important than genre and size. In Gamallo and González López (2010), the authors also relied on topic restrictions and language to gather comparable articles from Wikipedia. A complete different approach is described in Saralegi et al. (2008). In this work, the authors proposed to measure the comparability of a corpus by computing the *semantic similarities* at the document level. The hypothesis behind this is that the containment of many document pairs with a fairly high semantic similarity would improve terminology extraction based on context similarity. Braschler and Scäuble (1998) took advantage of external indicators to find similarities between pairs of documents. As the same story is usually published on similar dates by news agencies, they used the publication date as an indicator to align pairs of articles (Braschler and Scäuble, 1998:185). Yet, there are other features that could be considered. When documents are extracted from the Web, the structure and the context that describes the documents origin could be retrieved to classify them. An easy way to access this information is to look at the internal *HTML structure* marked by HMTL tags and analyse it using, for instance regular expressions (Goeuriot et al., 2009:57). By a way of example, Goeuriot et al., 2009:56 stated that, apart from the period, the document *authorship* could be used, since authors sharing the same style are likely to produce similar texts. In fact most of these works combine both *linguistic* and *extra-linguistic* criteria to compile and assess comparability content (e.g. Braschler and Scäuble, 1998; Goeuriot et al., 2009). Another example can be found in Bekavac et al. (2004) and Skadiņa et al. (2010b), in which the authors choose as parameters of comparability the domain and the topic as linguistic criteria and the size and the time span as extra-linguistic criteria. In the same line, Talvensaari et al. (2007) and Hashemi et al. (2010) used the document topics and their publication dates to align comparable documents.

In short, comparability is ensured by using several characteristics which can refer to the text creation context (publication dates, authorship, etc.), or to the text itself (topic, genre, etc.). Table 3 intends to put these features in perspective, i.e. tries to give a general idea, not only about the most common features used to measure the documents comparability, but also the most frequent retrieve mechanism used to access them.

	Similarity Features	**Retrieve Mechanism**
Linguistic	genre	words-frequency; keyword extraction; POS tagging; semantic similarity measures
	domain	
	type of discourse	
	topic	
Extra-linguistic	publication dates	regular expressions
	authorship	
	time span	
	size	
	HTML structure	

Table 3: Common similarity features used to find comparable content and measure the documents similarity along with the most common retrieving mechanisms.

2.4.2 Assessing Comparability and Parallelism

Once the appropriated features are correctly selected, the next step is to define heuristics to assess the internal corpus comparability degree. As mentioned in the previous section, one way to characterise a comparable corpus is through the degree of comparability that its documents share between each other. In theory, this may seem to be fairly straightforward, but in practice there are various factors to consider and consequently different criteria can be used to measure the degree of comparability. In detail, comparability and/or parallelism is considered a complex issue because there are different levels (e.g. document collections, individual documents, paragraphs and sentences) and features to consider (e.g. linguistic and extra-linguist). So far, there has been no agreement on the degree of similarity that documents in comparable corpora should have, or even agreement about the criteria for measuring parallelism and/or comparability. As pointed out by Sharoff, 2010:1, "the notion of comparable corpora rests on our ability to assess the difference between corpora which are claimed to be comparable, but this activity is still art rather than proper science". Nevertheless, there have been some attempts to determine and specify different levels of comparability/parallelism in comparable corpora (cf. Braschler and Scäuble, 1998; Bekavac et al., 2004; Fung and Cheung, 2004; Skadiņa et al., 2010a). In the next paragraphs these attempts are described in detail.

In Braschler and Scäuble (1998), the authors propose a five-level relevance scale in order to assess the quality of comparable documents alignment. The levels of relevance used to align pairs of documents are the following (Braschler and Scäuble, 1998:190):

i) **Same story**: where two documents cover exactly the same story/event.

ii) **Related story**: two documents deal with the same event or topic from a slightly different viewpoint. Alternatively, one of the documents may cover the same event or topic, but the topic is only a part of a broader story, or the article is composed by multiple stories.

iii) **Shared aspect**: two documents address various topics, but at least one of them is shared.

iv) **Common terminology**: the events or topics are not directly related, but they share a considerable amount of terminology.

v) **Unrelated**: the similarity between the documents is slight or non-existent.

Later, in 2004 Bekavac et al. introduced the notion of two levels of comparability of corpora (Bekavac et al., 2004). According to the authors, these levels of comparability could be called *light* and *hard* (Bekavac et al., 2004:1188):

i) **Light**: corpora are said to be lightly comparable when their similarity is only in terms of extra-linguistic and extra-textual features, such as size, time-span, text genres, gender and/or age of the authors, etc.

ii) **Hard**: hard comparable corpora is dependent on the previous collected lightly comparable corpora. In detail, this second type of comparability derives from

the first one by applying certain language technology tools/techniques[41] and some pre-defined parameters of their usage, with the purpose of finding out which documents in lightly comparable corpora deal with similar topics. Then, the resulted subsets of lightly comparable corpora that have been selected by those tools/techniques can be considered as "hard" comparable corpora.

Also in 2004, Fung and Cheung (2004) proposed three levels of comparability:

i) **Parallel**: a sentence-aligned corpus containing bilingual translations of the same document.

ii) **Noisy-parallel**: also called a "comparable" corpus, containing non-aligned sentences that are mostly bilingual translations of the same document, focused on the same thematic topics, with some insertions and deletions of paragraphs.

iii) **Very-non-comparable**: a corpus that contains far more disparate, very-non-parallel bilingual documents that could either be on the same topic or not (i.e. in-topic and off-topic, respectively).

Probably the most well-known work on this topic is presented in Skadiņa et al. (2010a), in which the authors present four levels of comparability of comparable corpora:

i) **Parallel**: texts considered as accurate or approximate translations with minor variations in language. They give examples of legal documents, software manuals, and fiction translations.

ii) **Strongly comparable**: texts closely related containing the same event or describing the same subject. The given examples are: texts written by the same source, with the same editorial control, in different languages; and texts concerning the same subject, written by independent news agencies (e.g. Wikipedia articles).

iii) **Weakly comparable**: texts of the same narrow or broader domain and genre, but describing different events, or varying in subdomains and specific genres. An example of that is the database administrator guide for MySql in two different languages.

iv) **Non-comparable**: pairs of texts that do not have much in common. The Web is an example of this type of texts.

Despite the concept of comparability is still considered a complex issue, several levels of comparability have been proposed so far (cf. Braschler and Scäuble, 1998; Bekavac et al., 2004; Fung and Cheung, 2004; Skadiņa et al., 2010a). Even though these four different approaches are not able to be directly compared (e.g. due to its subjectivity), table 4 put them side-by-side in order to show how do they correlate between each other.

As table 4 shows, the comparable corpora criteria defined by Skadiņa et al. (2010a) as "strongly" and "weakly" match to the "noisy-parallel" and "hardly"

[41] These techniques could be: simple comparison of frequency lists of lemmas and/or collocations; named entity recognition, classification and comparison; document classification; term extraction comparison; etc. (Bekavac et al., 2004:1188).

<table>
<tr><th></th><th>Braschler and Scäuble (1998)</th><th>Bekavac et al. (2004)</th><th>Fung and Cheung (2004)</th><th>Skadiņa et al. (2010a)</th></tr>
<tr><td rowspan="5">Linguistic Criteria</td><td>-</td><td>-</td><td>Parallel</td><td>Parallel</td></tr>
<tr><td>Same story</td><td rowspan="4">Hard</td><td rowspan="4">Noisy-Parallel</td><td>Strongly comparable</td></tr>
<tr><td>Related story</td><td rowspan="3">Weakly comparable</td></tr>
<tr><td>Shared aspects</td></tr>
<tr><td>Common terminology</td></tr>
<tr><td>Extra-linguistic Criteria</td><td>Unrelated</td><td>Light</td><td>Very-non-comparable</td><td>Non-comparable</td></tr>
</table>

Table 4: Levels of comparability in comparable corpora presented in the literature.

criteria, since they share the same and/or similar topic, as "hardly" and "noisy-parallel" criteria do (Bekavac et al., 2004; Fung and Cheung, 2004). Moreover, the first four levels defined by Braschler and Scäuble, 1998 also fall within this category. If, by on one hand Braschler and Scäuble (1998)'s classification presents a greater granularity, one the other hand Bekavac et al. (2004), Fung and Cheung (2004) and Skadiņa et al. (2010a) do not make any distinction between the information specificity shared between two documents. At the lower level of comparability, Braschler and Scäuble (1998) and Skadiņa et al. (2010a) do not explicitly consider any kind of extra-linguistic features in their criteria as Bekavac et al. (2004) and Fung and Cheung (2004) do. Finally, it is worth to notice that Fung and Cheung (2004) and Skadiņa et al. (2010a) reclaim a higher level of comparability, the parallel level, which corresponds to pairs of texts with minor variations in language.

2.5 Summary

In this section, we bridge the theoretical work described in the previous sections with the work developed in the scope of this thesis. The first part targets the compilation guidelines followed in this work and the second part briefly describes the techniques applied to assess and measure the internal degree of comparability in comparable corpora.

2.5.1 Comparable Corpora Compilation

Through this section, various corpora compilation techniques and tools were analysed in order to identify their limitations and propose new ways of improvement. Firstly, we started by defining the concept of comparable corpora (section 2.1). Then, the importance of their design/classification to the optimal results during their usage were presented in section 2.2. Finally, section 2.3 presented the five compilation protocol phases. These three sections are of paramount importance to one of the main contribution of this work, the iCompileCorpora application. The scarce number of tools available, the plethora of compilation protocols and performing issues associated with the current compilation tools on the market are just an example of the challenges that language users face when they try to integrate these tools in their daily workflow. This is largely due to the fact that in many cases the real needs of language users were not considered when designing these tools. An an attempt to fulfil this gap, the purpose of iCompileCorpora is to improve existing comparable corpora compilation tools by addressing their well-known shortcomings via the use of more sophisticated technologies along with better consideration of user requirements and feedback. In addition, this new web-based prototype not only aims at improving the compilation process but also allow quick development for new language pairs and, consequently improve the user satisfaction.

2.5.2 Assessing Comparable Corpora

In this work, we described and explored various techniques to assess and measure the internal degree of comparability in comparable corpora. After a careful analysis of the various ideas presented in the literature on how to classify comparable corpora (section 2.4), we identified that several variables and criteria are usually followed when building a corpus and assessing its level of comparability. In fact, current methods rely either on the pre-defined set of features used to compile the corpus or on the human analysis afterwards, yet little has been said about textual distributional similarity in this context and the quality that it brings to research. In an attempt to fulfil this gap, this work intends to present a simple but efficient methodology capable of not only measuring a corpus internal degree of relatedness, but also to increase it by helping the user to identify and filter out irrelevant document to the corpus. To do so, this methodology takes advantage of both various NLP technology and statistical methods in a attempt to automatically access the relatedness degree between sentences and documents.

Chapter 3

Compiling Corpora from the Web

"Everything should be as simple as possible,
but not simpler."

—Albert Einstein

This chapter summarises the research reported in Costa et al., 2014c; 2015d and Costa et al., 2015e. Each publication explores different aspects of the first Research Question (RQ1) discussed in section 1.2. Apart from these publications, this chapter presents and describes two tools deployed during this work, the SCleaner[42] (a web-based program that helps users formatting text) and the iCompileCorpora[43] (a multilingual web-based comparable corpora compilation tool).

In the last decade, there has been a growing interest in bilingual and multilingual corpora. In translation, in particular, their benefits have been demonstrated by several authors (cf. Bowker and Pearson, 2002; Bowker, 2002; Zanettin et al., 2003; Seghiri, 2015; Corpas Pastor, 2008; Corpas Pastor and Seghiri, 2009; 2016; Seghiri, 2016; 2017b). Their objectivity, reusability, multiplicity and applicability of uses, easy handling and quick access to large volumes of data are just some examples of their advantages. Thus, apart from manual, automatic and assisted translation, it is not surprising that the use of corpora has been considered an essential resource in several other research domains such as stylistics, terminology and language teaching. Ideally, parallel data would be the best resource both for language engineering such as Natural Language Processing (NLP) applications and for language users such as translators, interpreters or language learners (Cencini, 2002; Kotani and Yoshimi, 2015; Laviosa, 2016). Nevertheless, the lack of sufficient and up-to-date parallel corpora and linguistic resources, particularly for poorly-resourced languages and narrow domains is currently one of the major obstacles to further advancement in these areas. One potential solution is the exploitation of non-parallel bilingual and multilingual text resources, also known as comparable corpora, in other words, corpora that include similar types of original texts in one or more language using the same design criteria (cf. EAGLES, 1996b; Corpas Pastor, 2001). Although comparable corpora can compensate for the shortage of linguistic resources and ultimately improve both manual and automated translation quality, the problem of data collection is still a significant technical challenge. Existing solutions to compile comparable corpora are sometimes scarce, proprietary, simplistic with limited features or too complex to be used by laypersons. Accordingly, the main focus of this chapter is two-fold. Firstly identify the shortcomings and strengths of the current compilation tools available on the market. Secondly, with the aim of tackling their performance, usability and design problems, present an innovative multilingual web-based comparable corpora compilation prototype, which we named iCompileCorpora.

This chapter starts by describing and performing a careful analysis of the shortcomings and strengths of the most well-known comparable compilation tools available on the market (section 3.1). Then, section 3.2 illustrates how to use iCompileCorpora to build multilingual comparable corpora. Section 3.3, describes the SCleaner, a web-based tool built to help users formatting text. Finally, section 3.4 presents our main achievements and ideas for further improvements.

[42] http://www.lexytrad.es/scleaner/index.php
[43] https://icompilecorpora.herokuapp.com/home

3.1 Existing Comparable Corpora Compilation Solutions

Due to the fact that parallel corpora remain a scarce resource for poorly-resourced languages and often restricted to specific domains (e.g., political speeches, legal texts, news, etc.), the need for tools to build comparable corpora has increased (cf. Seghiri, 2017b). As a result, there is a growing literature on using the Web for constructing various types of text collections, including domain-specific monolingual, bilingual and multilingual comparable corpora (cf. Baroni and Bernardini, 2004; Baroni et al., 2006; Costa et al., 2014c; 2015d;e). Particularly, for translation purposes their benefits have been already demonstrated by several authors (cf. Bowker and Pearson, 2002; Bowker, 2002; Zanettin et al., 2003; Corpas Pastor and Seghiri, 2009; Seghiri, 2017b).

The Internet is a gold mine for documents in different languages covering overlapping information without being translations of each other. Nevertheless the process of retrieving comparable documents is not trivial. Although the process of compiling comparable corpora can be manually performed, nowadays, specialised tools can be used to automate this tedious task.

This section presents and describes in detail how the two most known tools on the market exploit corpora mined from the Web, as well as their advantages and drawbacks.

3.1.1 BootCaT

BootCaT[44] (Baroni and Bernardini, 2004) is a semi-automatic compilation program that makes use of online information to construct a Web-based corpus. The process is very simple and only requires a set of seed terms as input. Then, these seeds are randomly grouped to form tuples (i.e. a variety of combinations of the seeds), which are submitted as search query strings to the Bing[45] search engine API. BootCaT allows the user to define, before starting the retrieval process, a list of "black" and "white" words. If on one hand, the blacklist is used to exclude from the dataset documents, i.e., not include documents containing more than a certain number of words from the blacklist. On the other hand, the whitelist is used to make sure that documents above a certain threshold are included in the dataset. In other words, above a certain ratio between the words in the whitelist and the total number of words in the document. Then, during the download process, the top n pages returned for each query are retrieved and formatted as plain text.

As a result, BootCaT allows the user to retrieve a large amount of documents in just a few minutes, reducing the time of manual intervention in the compilation process. It is also possible to build a larger corpus by repeating the process using more seeds, or even create comparable corpora by repeating the process using similar keywords in different languages. Having this in mind, this tool is not based on automatic but semi-automatic search.

Despite of the multiple advantages, BootCaT has a few limitations, which constrains the compilation of mono-, bi- and multilingual comparable corpora.

[44] `http://bootcat.sslmit.unibo.it`
[45] `http://www.bing.com`

The following paragraphs summarise some of the them (Baroni and Bernardini, 2004:1313 and Gutiérrez Florido et al., 2013:3).

- ◇ lack of technical support, apart from the FAQs section there is no technical documentation available;
- ◇ the searches performed in the Web only uses the Boolean operator "AND", which consequently leads to less accurate searches than if Boolean operators such as "NOT" and "NOR" were used;
- ◇ in order to obtain results the seed words need to be semantically related to each other, if not, the retrieved documents will not have an acceptable quality;
- ◇ despite the possibility of choosing the lengths of the tuples, the tool restricts the possible combinations of the tuple's length, i.e. it is not possible to combine tuples with length of two and three at same time, for example;
- ◇ as reported by Gutiérrez Florido et al., 2013:3, sometimes the tool freezes during the search process, which may be due to: poor selection of keywords; a poor choice of URLs; the limit of searches per month[46]; or even due to internal problems;
- ◇ finally, the tool does not allow to perform a new compilation without closing and opening the tool again.

Despite some drawbacks, this tool can be seen as a viable source of "disposal" corpora (Varantola, 2003) built virtually for several purposes, such as translation tasks, construction of terminologies databases and domain-specific Machine Learning tasks (Baroni and Bernardini, 2004).

BootCaT toolkit can either be used in the form of a library (a suite of Perl scripts) or used as a graphical interface, i.e. a wizard that guides the user through the process of creating a Web corpus. It is important to mention that the interface is not as complete in terms of features as the command-line scripts. BootCaT is free and open source. In detail, the BootCaT front-end is a free software, developed in Java, that can be redistribute and/or modified under the terms of the GNU General Public License[47]. Regarding the BootCaT command-line scripts suite, it can be copied or redistributed under the same terms as Perl[48].

3.1.2 WebBootCaT

Sketch Engine[49] (Kilgarriff et al., 2004) is a leading corpus query tool. Apart from offering a corpus-building tool, it also provides access to corpora online and several analysis tools in a single platform. Nevertheless, the most relevant tool for this work is the WebBootCaT[50] (Baroni et al., 2006), which allows the user to create a

[46] BootCaT uses the Bing search engine to find web pages relevant to the domain. In order to perform this automated task BootCaT requires an account key from Bing, which has limits in the number of queries that can be submitted per month.

[47] `http://www.gnu.org/licenses/gpl.html`

[48] `http://dev.perl.org/licenses/artistic.html`

[49] `http://sketchengine.co.uk`

[50] `https://www.sketchengine.co.uk/webbootcat/`

specialised corpus from the Web in a few minutes. To do that, it only requires a set of seed words or URLs as input. This tool can be seen as a Web-service version of the BootCaT tool, but rather than download and install a software, WebBootCaT has the advantage of been already installed on a Web server. Yet, this tool is only freely available on a trial basis or through the commercial product subscription.

3.2 Towards a new Web-based Comparable Corpora Tool

The World Wide Web has become a primary meeting place for information and recreation, for communication and e-commerce. Millions of users have created billions of web pages in which they expressed their views about the world. As a source of machine-readable texts for corpus linguists and researchers in related fields such as Natural Language Processing (NLP) and Machine Translation (MT), the Web offers extraordinary accessibility, quantity, variety and cost-effectiveness textual data. This linguistic and cultural content is considered a gold mine for lexicographers, linguists, translators, teachers and other language professionals.

As a result, several tools, such as web crawlers, language identifiers, HTML parsers, HTML cleaners, etc. have been developed and combined in order to retrieve either general purposes or domain specific corpora from this gold mine. Nevertheless, the applicability of current data-driven methods directly depends on the availability of large quantities of parallel and or comparable data. By way of example, in the translation field, the translation quality of current data-driven MT systems varies dramatically from quite good, for language pairs with large corpora available (e.g. English and Spanish), to fairly unusable for under-resourced languages and narrow domains where little data is available (e.g. Croatian and Portuguese). Indeed, the majority of the European languages are under-resourced and lack parallel corpora or even language technologies for translation (Eisele and Xu, 2010). Even though comparable corpora can compensate for the shortage of linguistic resources and ultimately improve automated translations quality for under-resourced languages and narrow domains (Munteanu and Marcu, 2005; Eisele and Xu, 2010; Skadiņa et al., 2010a), the problem of data collection presupposes a significant technical challenge (Maia, 2003).

Although the compilation process could be manually performed, nowadays specialised tools can be used to automate this tedious task (Baroni and Bernardini, 2004; Baroni et al., 2006; de Groc, 2011). Nevertheless, as we described in the previous section (see section 3.1), these compilation tools are scarce or proprietary, simplistic with limited features and designed to compile one monolingual corpus at a time, in other words they do not completely fulfil the user's needs (Costa et al., 2014c). Consequently, their simplicity, lack of features, performance issues and usability problems (Gutiérrez Florido et al., 2013) result in a pressing need of improvement or even to design new compilation tools tailored to fulfil not only translators' and interpreters' needs, but also professionals' and ordinary people's (Costa et al., 2014c; 2015d;e).

Accordingly, this section aims at describing an web-based comparable corpora compilation prototype build with the purpose of solving some of the drawbacks on the current tools available on the market.

3.2.1 iCompileCorpora

In this section we present iCompileCorpora, a web-based compilation prototype capable of semi-automatically compiling monolingual and multilingual comparable corpora from the Web. For those familiar with BootCaT and WebBootCaT, the iCompileCorpora workflow will result very familiar. Despite some similarity with these two tools, iCompileCorpora aims at being more user-friendly, intuitive, flexible and robust. By a way of example, it was designed to build multilingual comparable corpora and make full usage of various Boolean operators, amongst other non-visible improvements goals such as performance and document formatting. Hereafter, we show how intuitive is the compilation process of a multilingual comparable corpora.

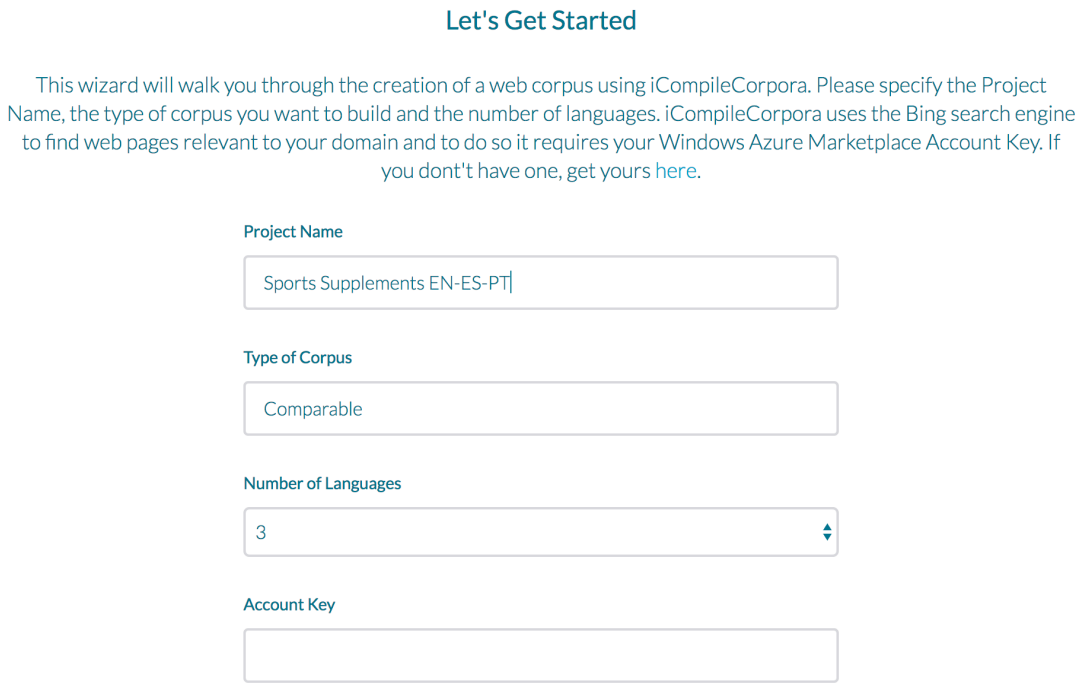


Figure 1: iCompileCorpora - Let's Get Started.

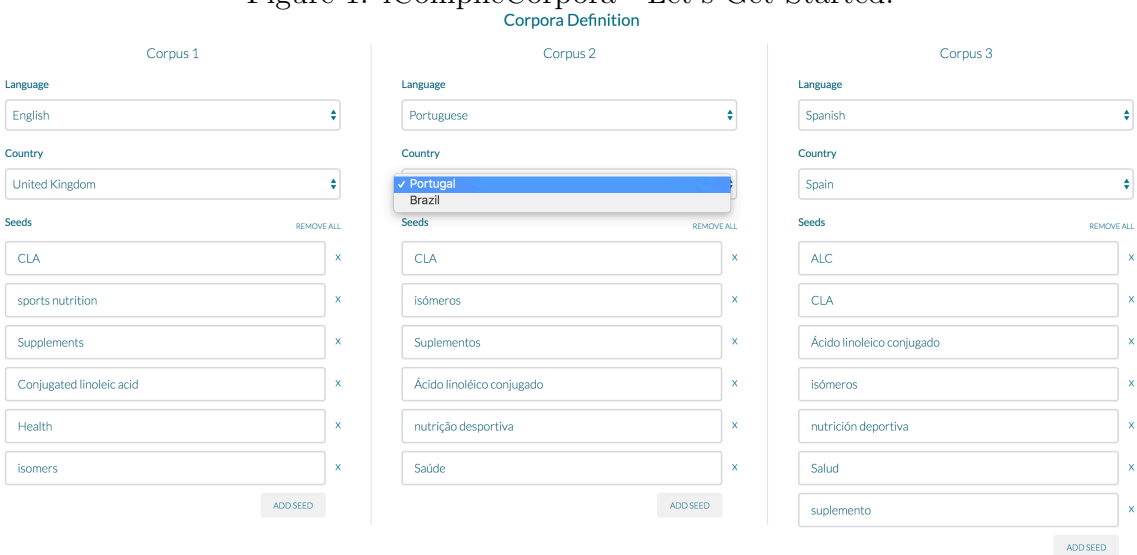


Figure 2: iCompileCorpora - Corpora Definition.

Figure 1 presents the "Let's Get Started" screen. This screen allows the user to define the "Project Name", the "Type of corpus"[51], the "Number of Languages" (minimum 1 and maximum 5) and to insert the user's Bing Search API "Account Key", which can be requested on the Microsoft Azure website[52].

In the next screen, Figure 2, "Corpora Definition", the user is requested to select the "Language", "Country" and the "Seeds", also known as keywords. The languages and countries available are dictated by the Bing API. Currently there are eighteenth languages available and the options on the field country depend on the selected language. By a way of example, for Portuguese the countries available are Portugal and Brazil. Which regards the "Seeds", the idea is to insert one seed per text box, which can be either a single or multiple words, and there is no limit in terms of number of seeds.

As soon as the user defines the seeds for the various corpora and clicks "Next", these seeds are randomly grouped to form queries, see figure 3. This step is as important as identifying the right seeds because these queries will be submitted to the search engine, and consequently the retrieved documents depend on these queries. Accordingly, it is important to make sure that the automatic generated queries make sense for the task in hand. If not, the user should manually edit, delete or add new seeds using the button "ADD SEED" (see figure 4). It is important to mention that it is possible to use all the three Boolean operators allowed by the Bing API, i.e."AND", "OR" and "NOT" to build the queries, and there is no restriction in terms of the length of the queries.

The next step in the process is to set-up the "Search Restrictions". Figure 5 presents the various options available: limit the search to specific domains, exclude specific domains, apply filters and select the desired number of results per query. The number of results per query will directly restrict the maximum of documents retrieved. For example, if selected 10, the maximum number of possible retrieved documents will be 10 times the number of queries. Please not that increasing the number of results per query will result in a larger corpus, yet its contents will tend to become less relevant.

Next, the tool starts collecting the URLs from the search engine (see figure 6). This process might be a bit slow because it depends on various factors, such as the number of queries, number of results per query, as well as on other external factors such as user's Internet connection speed and Bing's servers traffic. The lower text area shows in real time the URLs that are being collected from the search engine.

As soon as the tool receives all the URLs from the search engine the user can choose to remove URLs from the list that might not be interesting for the task in hand (see figure 7). Please note that it is possible to click on the URLs to visit the web page and decide whether it should be include in the corpus or not.

In the next step the tool tries to stablish connection with all the URLs and retrieve their content (see figure 8). It can happen that the number of retrieved documents is lower than expected. This happens when URLs are no longer accessible (the page is no longer online or its content is protected, etc.). When this happens the tool highlights them in red (see the second URL in "Corpus 3", figure 8).

During the crawling process, the tool automatically removes HTML tags, formats

[51] Only the option Comparable Corpora is available at the moment. In the future the idea is to also allow the user to use a similar pipeline to build Parallel Corpora.

[52] `https://azure.microsoft.com/en-us/try/cognitive-services/`

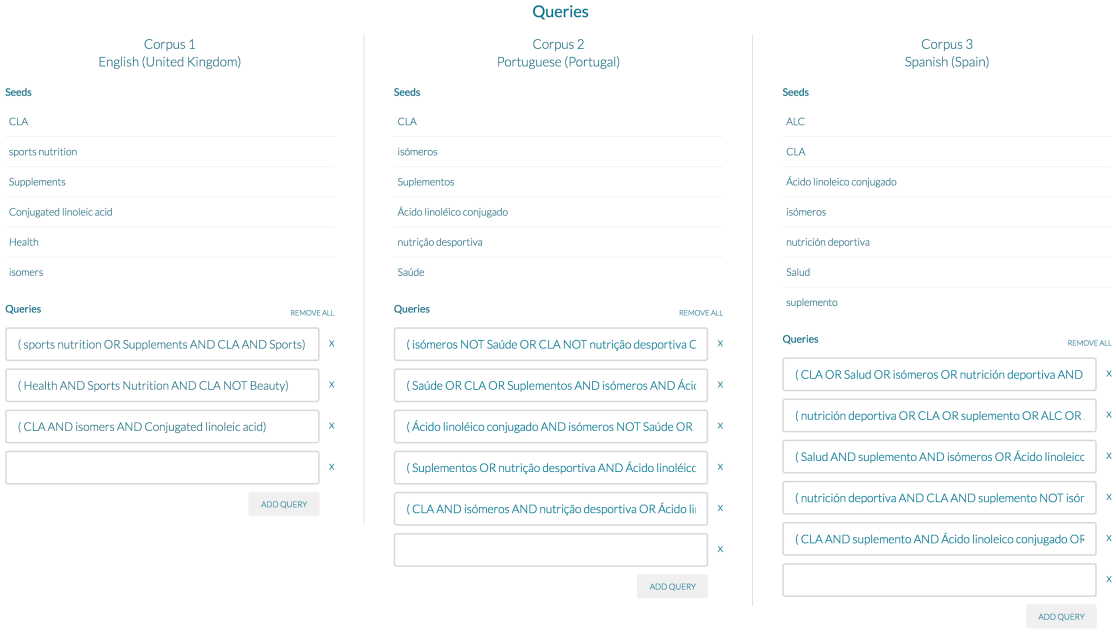


Figure 3: iCompileCorpora - Setting up the Queries (part 1).

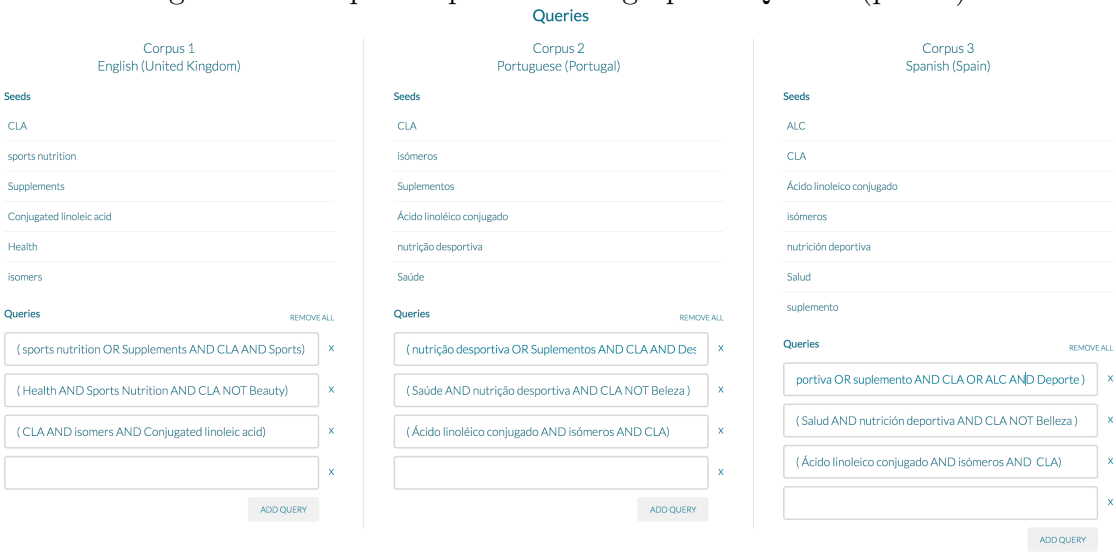


Figure 4: iCompileCorpora - Setting up the Queries (part 2).

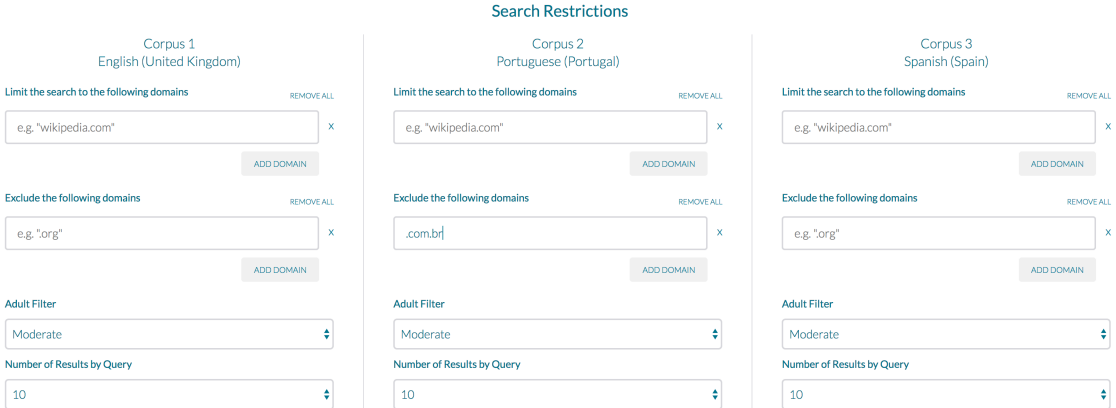


Figure 5: iCompileCorpora - Search Restrictions.

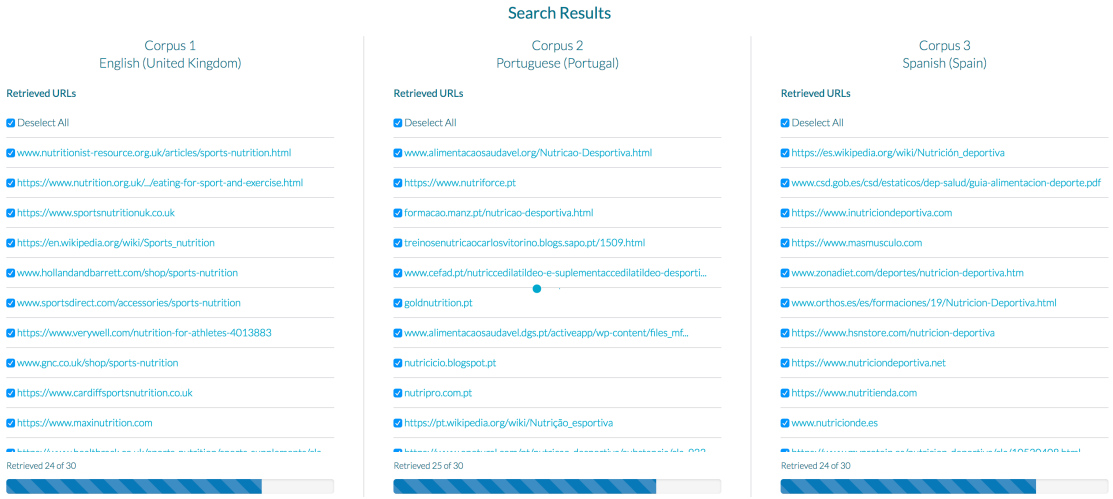


Figure 6: iCompileCorpora - Search Results - Retrieving URLs.

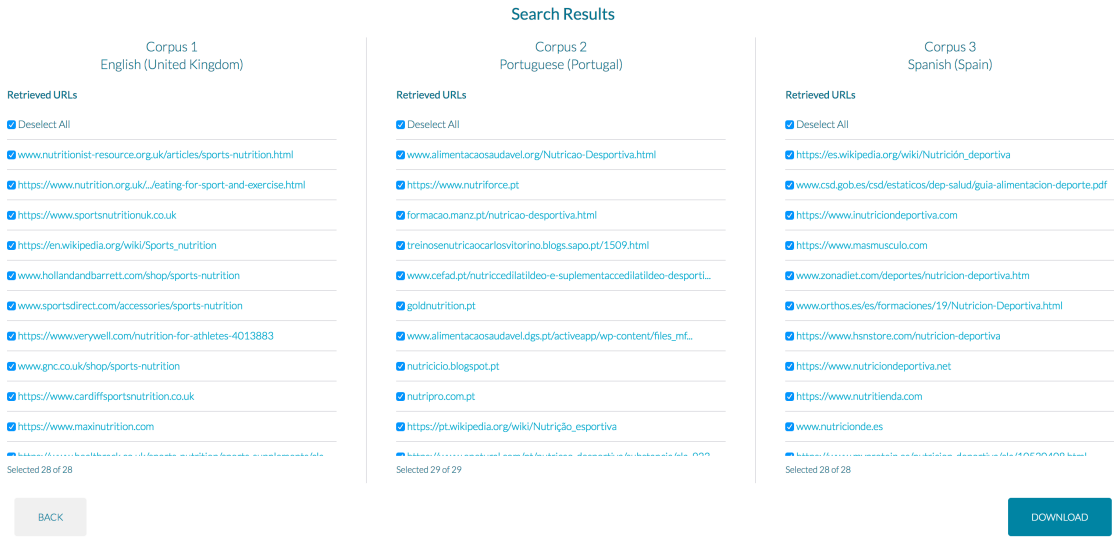


Figure 7: iCompileCorpora - Search Results - Select Retrieved URLs.

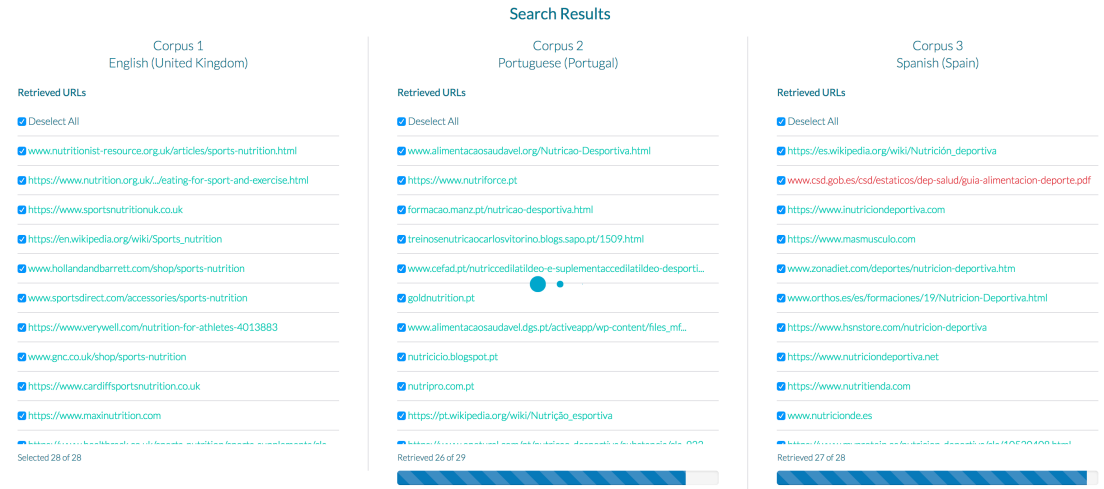


Figure 8: iCompileCorpora - Search Results - Downloading Documents.

the text and converts it to plain text. When the process is finished, the tool compresses all the documents in a zip file and prompts a dialogue box to save the file locally. Finally, the user can start working on the corpora by exploring the various subcorpora and associated documents (see figure 9).

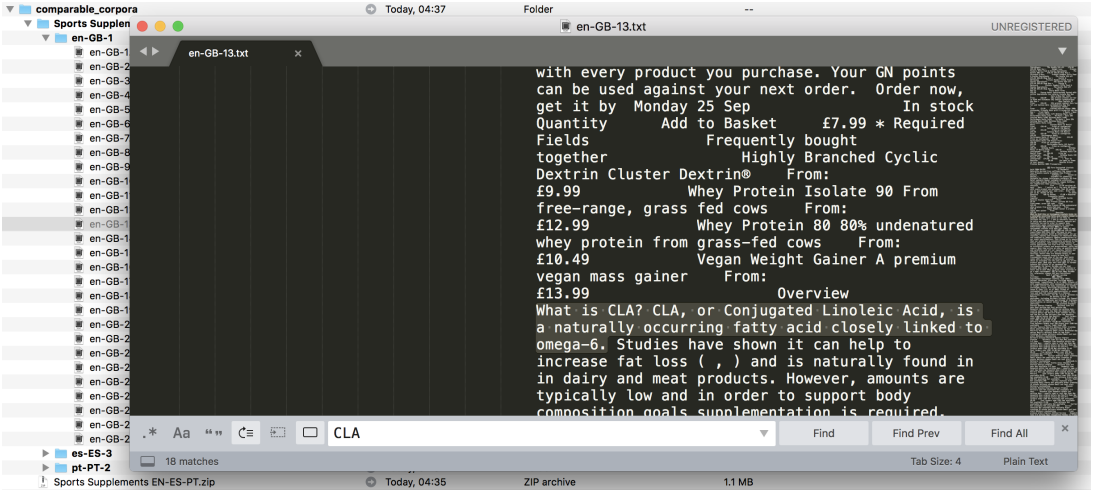


Figure 9: iCompileCorpora - The "Sports Supplements EN-ES-PT" Corpora.

3.2.2 Final Remarks and Directions

This section aims at comparing the comparable corpora compilation prototype developed during the period of this research with those already on the market, as well as pointing out directions and improvements that could be made to improve the compilation process.

Although the aforementioned tools can not be directly compared –mostly because of external factors, such as the constant changes on the search APIs used by these tools and the unknown parameters used to retrieve the documents–, it is easy to spot some similarities in terms of features offered during the compilation process. Accordingly, table 5 aims at putting them in perspective so we can have a better understanding on how do they either overlap or differ from each other in terms of features.

Hereafter we focus on discussing the most distinctive features. Starting by the "Availability", both BootCaT and iCompileCorpora are open source tools that can be freely used and forked from their respective web repositories. On the other hand, WebBootCaT is part of the giant Sketch Engine's ecosystem, which can only be used under a monthly subscription. Nevertheless, the subscription not only grants access to the comparable compilation tool, but also to other tools such as keyword extraction or parallel concordance, amongst other corpora management tools. Other than the subscription, the only other difference between BootCaT and WebBootCaT is the "Type of Application". BootCaT can be used either as a standalone or library, and WebBootCaT only as a web-based service. From table 5 we can observe that the biggest different between iCompileCorpora and the other two tools is the non-existent option to upload a list of white or black words. Yet, it should not be seen as a huge disadvantage because iCompileCorpora offers the freedom to edit and use multiple Boolean operators to guarantee optimal output

results ("Boolean Operator(s)" and "Queries Manipulation").

Feature	BootCaT	WebBootCaT	iCompileCorpora
Availability	free	proprietary with demo	free
Type of Application	standalone & library	web-based	web-based
Black List	yes	yes	no
White List	yes	yes	no
Boolean Operator(s)	AND	AND	AND & OR & NOT
Queries Manipulation	restricted	restricted	free
Corpora	monolingual	monolingual	multilingual
Search Engine	Bing API	Bing API	Bing API
Output format	plain text	plain text	plain text

Table 5: Comparable Compilation Tools: *BootCaT*, *WebBootCaT* and *iCompileCorpora*.

Despite some similarity between these three tools, iCompileCorpora was built with the main purpose of being user-friendly, intuitive, flexible and robust. For example, it allows the user to build multilingual comparable corpora at a time and make full usage of various Boolean operators while creating the searchable queries, amongst other non-visible improvements than meets the eye, such as performance, document formatting and user experience. Although more testing would be required to be considered a stable tool, we can state that the main purpose has been achieved by building a simple compilation interface with simplified steps for both novices and advanced users.

To sum up, we believe that we made a step in the right direction by showing that is possible to take advantage of the current technologies and build a simple, yet robust multilingual comparable compilation tool that is intuitive and easy-to-use by both professionals and laypersons. However, there is still much room for improvement. We see this tool as a futurist multi-purpose tool that can be used not only to build comparable corpora but also parallel corpora. Moreover, other features can be explored in the future (see for example the ideas described in Costa et al., 2015e). By a way of example, a set of concordance features, such as search for words in context, automatic extraction of the most frequent words and multi-words, or even management features such as edit, copy and paste sentences and documents from and to documents and corpora respectively, as well as manage corpora into domains and sub-domains could be interesting add-ons to be incorporated. Nevertheless, the obvious next step would be to continuously implement and test one feature at time based on users' feedback, so we could finally come up with a compilation tool that fulfils everyone's expectations.

3.3 SCleaner

SCleaner[53] is a simple web-based tool created to help researches formatting unformatted documents.

While creating manual corpora or even when there is a need to add extra documents to existing corpora, which often are stored in a plain text format, researchers face the excruciating task of manually cleaning and formatting documents themselves. For example, when converting a PDF file to plain text or copying from a PDF and pasting into plain text, researchers can face

[53] `http://www.lexytrad.es/scleaner/`

various formatting problems such as white spaces, tabulations, sentence boundaries, amongst other formatting problems. Accordingly, SCleaner was build to help them with this boring and time-consuming task. In general terms, the tool uses various regular expressions to automatically remove extra tabs and white spaces and split sentences.

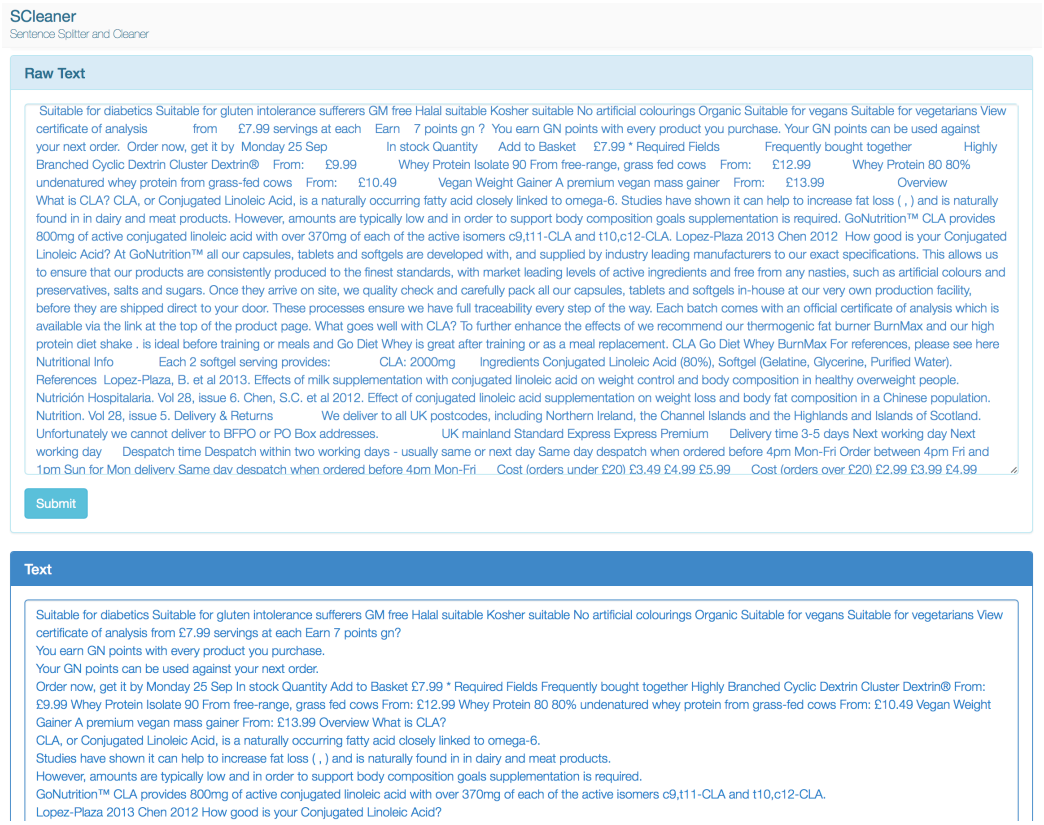


Figure 10: SCleaner Interface.

Figure 10 shows and example of what the tool can do. Firstly the user copies the text from an external source such as a PDF file, a web page, amongst other formats, and pastes its content in the text box "Raw Text" (see figure 10). Then, with a press of a bottom the text is analysed, cleaned and formatted and presented in a second text box "Text". Now the user just needs to copy the formatted text and continue with the task in hand.

SCleaner is a simple, yet very handy tool to have when dealing with unformatted text as it can speed-up this tedious and time-consuming task. It is important to mention that SCleaner is an open-source tool that can be used and forked from GitHub[54].

[54] `https://github.com/hpcosta/SCleaner`

3.4 Summary

This chapter summarised the research carried out in Costa et al., 2014c; 2015d and Costa et al., 2015e. Each section explored the first Research Question (RQ1), discussed in section 1.2 from different angles. In detail, this chapter analysed the shortcomings and strengths of current tools available on the market (section 3.1). Then, in section 3.2 was presented the iCompileCorpra, a multilingual web-based comparable corpora compilation prototype designed to increase the flexibility and robustness of the compilation process. Finally, a web-based program that helps users formatting text was introduced in section 3.3.

After a careful analysis of the most known comparable compilation tools on the market, several limitations and drawbacks were identified. Despite of the extraordinary effort and time invested on these tools, they are not keeping up to the current user's requirements, and the technology they are build on can be sometimes considered obsolete. In an attempt to show the research community that it is possible to fuse various disciplines, such as Software Engineering, User Experience (UX) and Computational Linguistics to deploy reliable software that tackles the current usability problems and performance issues found in the current tools on the market, a new comparable corpora tool named iCompileCorpora has been created. iCompileCorpora is the result of one person' three months' work, yet, it can be considered a reliable and intuitive peace of software that can be further improved due to its open-source policy.

To sum up, there is still a long road ahead of us before we can acknowledge and fulfil all the user's requirements. Nevertheless, we believe we made a step in the right direction by showing that is possible to take advantage of the current technologies and build a simple, yet robust multilingual comparable compilation tool that is easy and intuitive to be use by both professionals and laypersons. The next step would be to continuously improve and implement new features based on user's feedback, so we could finally come up with a comparable corpora compilation tool that fulfils everyone's expectations.

Chapter 4

Assessing Terminology Tools based on the Users' Requirements

"Don't be afraid of the space between your dreams and reality.
If you can dream it, you can make it so."

—Belva Davis

This chapter describes the studies reported in Costa et al., 2014b; 2015b; 2016b and Costa et al., 2017. Each publication explores different aspects of the second Research Question (RQ2) discussed in section 1.2.

Terminology tools have become an indispensable resource in education, research and business. Today, users can find a great variety of terminology tools of all kinds, and they all offer different features. Apart from many other areas, these tools are especially helpful in the professional interpreting and translation setting. We do not know, however, if the existing tools have all the necessary features for these kind of work and how to evaluate them according to interpreters' and translators' needs. Accordingly, the second main goal of this work is to address translators' and interpreters' needs and suggest new methodologies to help them increase the productivity and ease their labour-intensive activities, mostly in the preparation stage of a given task. To do so, firstly we identified the users' requirements regarding the use of terminology tools by analysing various users' surveys in the literature. Then, we established a set of well-defined and measurable features that permitted us to assess and distinguish various well-known terminology tools on the market concerning interpreters' and translators' needs in such a way that the results would be useful for both end-users as well as to the designers of such systems.

In detail, section 4.1 offers a tentative catalogue of technology-assisted interpreting tools, divided into Terminology Management Systems (TMS), note-taking applications for consecutive interpreting, voice recording applications and training tools. Then, section 4.2 highlights some of the features that interpreters expect from a TMS and proposes to standardise them into a discriminative scoring system so it could be used to evaluate current TMS available on the market. Section 4.3, focuses exclusively on Terminology Extraction Tools (TET) for translators with a view to identify the priorities for the design and features to be included in a TET. Then, a comparative analysis of various well-known TET currently available on the market based on the translators' most favourite features is made. It is important to mention that in the end of each section we report our main research findings and give some cues for further work. Finally, section 4.4 presents our main findings and highlights some ideas to improve current interpreters' and translators' terminology tools.

4.1 Technology-Assisted Interpreting: A Catalogue

This section summarises the work reported in Costa et al., 2014b. In detail, this article offers a tentative catalogue of current language technologies for interpreters, divided into Terminology Management Systems (TMS) for interpreters (which was further extended in Costa et al., 2017 and Corpas Pastor, 2018), note-taking applications for consecutive interpreting, voice recording applications and training tools.

In the last decades, several tools and applications have been created to meet the needs in different interpreting contexts and modes. Even though some interpreters still store information and terminology on scraps of paper or excel spreadsheets, there are some specialised computer and mobile software that can be used to compile, store, manage and retrieve information. They can typically be used to automate the process, increase the productivity and ease the labour-intensive activities of an interpreter before and during an interpretation service. Some of those applications are quite similar to the look-up terminology tools currently used by translators (Durán Muñoz, 2012). In fact, some of them have been developed to cater to the needs of both translators and interpreters. Hereafter we categorise these applications into note-taking (which help to minimising the processing effort), voice recording (which allow to organise, annotate and synchronise text, images, sound and video), computer-assisted interpreter training tools (which facilitate to practice audio or video clips interpreting exercises), and TMS (which permit to compile, store and search within glossaries).

Note-taking Applications: Consecutive interpreters use a specific system of taking notes to retrieve part of their source speech understanding from memory while minimising the processing effort. This supporting technique is usually performed with a traditional pen and paper. However, as more and more interpreters are turning to mobile devices to take notes, it is just natural that those devices become the favourite note-taking and ubiquitous capture tool on the go. Some examples of automated note-taking applications are Evernote[55], Inkeness[56] and Penultimate[57]. Along the same line, there is a computer-assisted tool for semi-automation of the note-taking in consecutive interpreting presented in Rafajlovska, 2013. This application provides a keyword with the most frequent symbols used by consecutive interpreters, which are linked to two *ad hoc* parallel dictionaries (Macedonian/English and Macedonian/French). By using the keyword, consecutive interpreters can take the same notes as they could on paper, but then they can also convert those notes into a readable message and save it for future reference. Today digital pens are capable of linking the written notes with ambient sound and upload it to a computer, allowing the interpreter to focus on listening and participating instead of worrying about catching every word during an event. Sky Wifi Smartpen, Echo Smartpen and Livescribe commercialised by Livescribe Inc.[58]

[55] https://evernote.com
[56] www.fenrir-inc.com
[57] http://evernote.com/penultimate
[58] www.livescribe.com

and Equil JOT[59] are just some examples of this amazing technology.

Voice Recording Applications: There are currently a number of applications that allow voice recording for training practice. GoodReader[60] and Documents[61], Audacity[62], Adobe Audition[63], AudioNote[64] are just some examples of applications capable of managing text, images, audio and video files.

Computer-assisted Interpreter Training Tools: Text-to-speech apps for iPad can also be successfully applied to teaching and improving language skills. For example, Voice Dream Reader[65], Voxdox[66] and Talk - Text to Voice[67] allow users to listen to words, texts, e-mail in several languages and formats. Regarding integrated tools capable of assisting interpreters during their services or when training, they are quite scarce. An exception is the Black Box (Bendazzoli and Sandrelli, 2009), a computer-assisted interpreter training tool designed to help interpreters work with a range of different materials (texts, audio, video, different types of exercises) and store their results for later review. It can be used to practice in different ways either by interpreting some audio or video clips or by doing some practical interpreting exercises, such as shadowing, cloze exercises or sight translation. It also allows teachers to edit and break down video and audio recordings to create different exercises and adapt authentic conference materials to the students' level of expertise. Black Box can be considered a suitable training workbench for trainee interpreters. Other web-based environments have recently been created along similar lines. InterpretaWeb[68] and Linkinterpreting[69] provide interpreters and students with a wide range of exercises (cloze, memory, cluster), and complete speeches to practice simultaneous and consecutive interpreting, along with information resources and news related to interpreting. These websites are of great use to students and for novice interpreters who are willing to practice and improve their interpreting skills.

Terminology Management Systems (TMS): There are some specialised computer and mobile software that can be used to quickly compile, store, manage and search within glossaries. Hereafter, we present some of the most outstanding applications developed by and for interpreters. They can be typically used to prepare an interpretation, in consecutive interpreting or in a booth.

Standalone software is probably the most popular type of software today (Costa et al., 2017), and TMS are no exception. Standalone TMS are tools that can be installed on the computer and operate independently of any other device or system.

[59] www.myequil.com
[60] www.goodiware.com
[61] http://readdle.com
[62] http://audacity.sourceforge.net
[63] www.adobe.com/products/audition.html
[64] http://luminantsoftware.com/iphone/audionote.html
[65] www.voicedream.com
[66] www.voxdox.net
[67] https://plus.google.com/communities/107986392540899459664
[68] www.interpretaweb.es
[69] http://linkterpreting.uvigo.es

Examples of standalone TMS are Intragloss[70], InterpretBank[71] and Interplex UE[72]. Intragloss is an intuitive and easy-to-use tool that facilitates the interpreters' terminology management process by producing glossaries (imported or created *ad hoc*), by searching on several websites simultaneously, by highlighting all the terms in the documents that appear in the domain glossary and by comparing different language versions of a document. However, it is currently platform dependent and only works on Mac OS X platforms. InterpretBank has a user-friendly, intuitive and easy-to-use interface, which allows us to import and export glossaries in different formats and suggests translation candidates by taking advantage of online translation portal services, such as Wikipedia, MyMemory and Bing. However, it is platform-dependent (it only works on Windows and Android), does not handle documents (only glossaries) and requires a commercial license. Interplex UE has a user-friendly interface and it is regularly updated. It allows us to import and export glossaries from and to various formats. However, it, too, is platform dependent (Windows and iOS only), does not handle documents, only glossaries, and requires a commercial license.

Broadly speaking, web-based TMS can be considered more sophisticated than standalone TMS since they include more advanced features and offer professional support, as they are mostly designed for commercial purposes. Although they were not built to help interpreters during the interpretation process, they can be extremely useful before the interpreting service as they allow them to store and share terminology more easily, especially for companies who have a considerable number of employers or freelance interpreters in a collaborative environment. Examples of most innovative and consequently more expensive web-based terminology management solutions on the market today are WebTerm[73], Acrolinx[74] and Termflow[75].

Mobile terminology apps are undoubtedly the next step in this ever-evolving domain of term management. TMS apps are systems which have been developed or optimised for small handheld devices, such as mobile phones, smartphones, iPads or PDAs, among others. Some of the most popular ones are Glossary Assistant[76] (a user-friendly multilingual glossary management application created by a professional team of interpreters for Android devices) and The Interpreter's Wizard[77] (a free iPad application capable of managing bilingual glossaries in a booth).

This section presented an overview of tools and applications available for interpreting practice and training. Although the number of these technologies is growing fast due to an increasing interest towards interpreters' needs, they are still insufficient and unable to fulfil all the necessary requirements. There is an urgent need to develop technologies that automate the process, increase the productivity and ease the labour-intensive activities of an interpreter before and during an interpretation service. Given that TMS are probably the most used assisting tools used by interpreters, there is a pressing need to help them to choose the tool that

[70] https://intragloss.com/
[71] www.interpretbank.de
[72] www.fourwillows.com
[73] www.star-group.net/en/products/webterm.html
[74] www.acrolinx.com/platform-services/terminology-management/
[75] www.termflow.de/
[76] http://swiss32.com
[77] http://the-interpreters-wizard.appsios.net/

best caters for their specific needs. Accordingly, the next section aims at establishing a set of specific and measurable features that permit interpreters to assess and distinguish the different tools concerning individual's and/or company's needs in such a way that the results would be useful for both potential customers as well as to the designers of such systems.

4.2 Assessing Terminology Management Systems for Interpreters

This section summarises the work reported in Costa et al., 2014a, Costa et al., 2015b and Costa et al., 2017, which aimed at describing and comparing current Terminology Management Systems (TMS) with a view to establishing a set of features in order to assess the extent to which terminology tools meet the specific needs of interpreters.

As in translation, domain-specific terminology becomes a cornerstone in interpreting when consistency and accuracy are at stake. Hence, an efficient use and management of terminology will enhance interpreting results. As a matter of fact, interpreters have limited time to prepare for new topics and they have to carry out searches and preparation prior to an interpretation and have it accessible during the interpreting service. Fortunately, there is an ever-growing number of applications capable of assisting interpreters before and during an interpretation service, even though they are still few if compared to those devoted to translators. Although these tools appear to be quite similar, they provide different kind of features which result in different degrees of usefulness. Accordingly, this section aims at shedding some light on a specific type of technology targeting interpreters, i.e. TMS and to carry out a comparative analysis of several of those tools in order to assess their relevance.

The remainder of this section is structured as follows. Section 4.2.1 highlights some of the features that interpreters expect from a TMS and proposes to standardise them into a discriminative scoring system so it could be used to evaluate current TMS available on the market. Then, section 4.2.2 presents and compares seventeen TMS with the aim of assessing them on the basis of a set of measurable features. Finally, section 4.2.3 presents the final remarks and highlights some ideas to improve current TMS.

4.2.1 Towards a Discriminative Scoring System

Although most of the current TMS on the market can be used to prepare a given interpretation, these systems differ from one another in their functionalities, practical issues, degrees of user-friendliness and target audience (i.e. individual or enterprise usage). Therefore, it is necessary to establish a set of specific and measurable features that permitted us to assess and distinguish the different tools concerning individual's and company's needs in such a way that the results would be useful for both potential customers as well as to the designers of such systems.

After a careful analysis of the priorities for the design and features to be included in a TMS reported in Moser-Mercer, 1992, Bergenholtz and Tarp, 2003, Tarp, 2008, Spohr, 2009, Rodríguez and Schnell, 2009 and Bilgen, 2011, we identified 15 main

features. Although some of them were identified as fundamental due to their extreme importance when assisting interpreters before and during an interpretation service, others are mostly related with the tools' design and surrounding. For instance, the "freedom to define the basic structure" identified in Rodríguez and Schnell, 2009 was reformulated into several practical measurable features, such as "N. of descriptive fields", "N. of working languages" and "N. of languages per glossary". Moreover, the possibility of "developing multilingual mini-databases", also identified in their study, was reconsidered as measurable features by means of the following criteria: "Manages multiple glossaries" and "N. of languages per glossary". Another example is the "Remote Glossary Exchange" measurable feature, which was inferred from the study conducted by Bilgen, 2009, who identified the need to exchange terminological information.

In an attempt to standardise these 15 features into a discriminative 0-100 scoring system, we used the EAGLES framework (EAGLES, 1996a) for the evaluation of NLP systems as a reference to divide these features into two categories: fundamental and secondary. Although all the reported characteristics are important to any software, in our work we mainly focused on the functionality of the software. Thus, we considered fundamental all the features related with the software's functionality ("A set of attributes that bear on the existence of a set of functions and their specified properties. The functions are those that satisfy stated or implied needs"), such as the management of multiple glossaries, the number of possible working languages permitted by the tool, how many of these languages can be used at the same time per glossary, the number of descriptive fields allowed per glossary entry and the possibility of managing terminology with preparation documents. The remaining 10 features, which are related with reliability, usability, efficiency, maintainability and portability where categorised as secondary. In detail, the features classified as fundamental to a terminology tool were given 10 points and 5 points to the secondary ones –except for web-based TMS, in which we removed one feature and considered 6 as fundamental and 8 as secondary. As we will see in the next section, we used these features to evaluate seventeen tools (9 standalone, 6 web-based and 2 mobile) and to assess which one of those was the most complete, both considering each sub-group separately and all the tools together.

4.2.2 Evaluating Terminology Management Systems

It is a well-known fact that terminology work is present in the whole process of preparation prior to an interpretation service. By a way of example, interpreters become familiar with the subject field by searching for specialised documents, by extracting terms and looking for synonyms and hyperonyms, by finding and developing acronyms and abbreviations and by compiling a glossary. According to Rodríguez and Schnell, 2009, interpreters tend to compile in-house glossaries tailored to their individual needs as the main way to prepare the terminology of a given interpretation. As previous studies and surveys have shown, this terminology management carried out by interpreters is frequently done manually or with very little help of technology.

However, in the last decade a wealth of TMS that interpreters could use to quickly compile, store, manage and search within glossaries have been developed. They can be typically used to prepare an interpretation, in consecutive interpreting or in a

booth. Even though most of these TMS have not been specifically developed for interpreters but for translators, there are some of them that cater for the needs of both translators and interpreters (Durán Muñoz, 2012; Costa et al., 2016b). Hereafter, we present some examples of standalone, web-based and mobile TMS.

Standalone TMS require an installation process and work as independent computer programs. Examples of standalone TMS are Intragloss[78], InterpretBank[79], Interplex UE[80], SDL MultiTerm Desktop[81], AnyLexic[82], Lingo[83], UniLex[84], TermX[85] and Terminus[86]. For more details about the aforementioned tool please see Costa et al., 2014a and Costa et al., 2017. Table 6 provides a comparative summary of the main features that characterise the TMS described in Costa et al., 2014a and Costa et al., 2017. Overall punctuations have been assigned for relevance and wealth of functionalities.

There also exist web-based tools, which work within a browser. Examples of most innovative and consequently more expensive web-based terminology management solutions on the market today are flashterm[87], Interpreters' Help[88], ASPLex[89], WebTerm[90], Termflow[91] and Acrolinx[92]. These tools can be considered more sophisticated than standalone TMS since they include more advanced features and offer professional support, as they were specially designed for commercial purposes. Although they were not built to help interpreters during the interpretation process, they can be extremely useful before the interpreting service as they allow them to store and share terminology more easily, especially for companies who have a considerable number of employers or for free-lance interpreters in a collaborative environment. Table 7 provides a comparative summary of the main features that characterise the web-based TMS mentioned above. As in previous cases, overall punctuations have also been offered to serve as a quick guide or checklist for interpreters.

[78] https://intragloss.com/
[79] www.interpretbank.de
[80] www.fourwillows.com
[81] www.sdl.com/cxc/language/terminology-management/multiterm/
[82] www.anylexic.com
[83] www.lexicool.com/soft_lingo2.asp
[84] www.acolada.de/unilex.htm
[85] www.translex.co.uk/software.html
[86] www.wintringham.ch/cgi/ayawp.pl?T=terminus
[87] www.flashterm.eu/home
[88] www.interpretershelp.com/
[89] www.termnet.nl/ASPLex.html
[90] www.star-group.net/en/products/webterm.html
[91] www.termflow.de/
[92] www.acrolinx.com/platform-services/terminology-management/

Feature	**SDL MultiTerm** 2014 (2013)	**TermX** (2013)	**Intragloss** 1 (2014)	**AnyLexic** 4 (2011)	**Lingo** 4 (2011)	**InterpretBank** 3.102 (2014)	**Terminus** 3.1 (2009)	**Intraplex** 2.1.1.47 (2012)	**UniLex** 0.9 (2007)
Manages multiple glossaries ⟨no=0; yes=10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	no ⟨0⟩
N. of possible working languages ⟨≤100=4; >100=7; unlimited=10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	180 ⟨7⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	35 ⟨4⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	30 ⟨4⟩
N. of languages per glossary allowed ⟨≤3=5; ≥4=10⟩	unlimited ⟨10⟩	6 ⟨10⟩	2 ⟨5⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	2 ⟨5⟩	5 ⟨10⟩	unlimited ⟨10⟩	2 ⟨5⟩
N. of descriptive fields ⟨non=0; 1=3; [2-5]=7; >5=10⟩	>5 ⟨10⟩	> 5⟨10⟩	4 ⟨7⟩	1 ⟨3⟩	>5 ⟨10⟩	4 ⟨7⟩	2 ⟨7⟩	non ⟨0⟩	2 ⟨7⟩
Handles documents ⟨no=0; yes=10⟩	no ⟨0⟩	no ⟨0⟩	yes ⟨100⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩
Unicode compatibility ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩
Imports from ⟨1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	MS Word, Excel & other CAT formats ⟨5⟩	MS Word, Excel & other CAT formats ⟨5⟩	MS Word, Excel & Plain Text ⟨3⟩	Excel, Plain Text & AEF ⟨3⟩	TMX & Plain Text ⟨2⟩	MS Word, Excel, TMEX & Plain Text ⟨4⟩	Excel & Plain Text ⟨2⟩	MS Word, Excel & Plain Text ⟨3⟩	Plain Text ⟨1⟩
Exports to ⟨non=0; 1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	MS Word, Excel & other CAT formats ⟨5⟩	MS Word, Excel & other CAT formats⟨5⟩	MS Word & Excel ⟨2⟩	Excel, Plain Text & AEF ⟨3⟩	TMX & Plain Text ⟨2⟩	MS Word, Excel, TMEX, Android & Plain Text ⟨4⟩	RTF, PDF & Plain Text ⟨3⟩	MS Word, Excel & Plain Text ⟨3⟩	Plain Text ⟨1⟩
Embedded online search for translation candidates ⟨no=0; yes=5⟩	no ⟨0⟩	no ⟨0⟩	yes ⟨5⟩	no ⟨0⟩	no ⟨0⟩	yes ⟨5⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩
Interface's supported languages ⟨1=1; [2-5]=3; >5=5⟩	> 5 ⟨5⟩	English ⟨1⟩	English ⟨1⟩	> 5 ⟨5⟩	English ⟨1⟩	English ⟨1⟩	English ⟨1⟩	English ⟨1⟩	English + 3 ⟨3⟩
Remote Glossary Exchange ⟨no=0; yes=5⟩	yes ⟨5⟩	no ⟨0⟩	no ⟨0⟩	yes ⟨5⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩
Well-documented ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩
Availability ⟨proprietary without demo=1; proprietary with demo=3; free=5⟩	proprietary without demo ⟨1⟩	proprietary without demo ⟨1⟩	proprietary without demo ⟨1⟩	proprietary with demo ⟨3⟩	proprietary with demo ⟨3⟩	proprietary with demo ⟨3⟩	proprietary with demo ⟨3⟩	proprietary with demo ⟨3⟩	free ⟨5⟩
Operating System(s) ⟨1=1; 2=3; ≥3=5⟩	Windows ⟨1⟩	Windows ⟨1⟩	Mac OS X ⟨1⟩	Windows ⟨1⟩	Windows ⟨1⟩	Windows & Android ⟨3⟩	Windows ⟨1⟩	Windows & iOS ⟨1⟩	Windows ⟨1⟩
Other relevant features ⟨subjective analysis=max. 5⟩	it is a concept oriented-tool and permits to add illustrations into each entry ⟨5⟩	availability to import and export from and to CAT tools ⟨5⟩	allows to highlight terms in the documents and merge a glossary with a document making it annotated to be printed ⟨5⟩	allows to share within a group of AnyLexic users ⟨1⟩	permits to add an unlimited number of descriptive fields ⟨5⟩	the MemoryMode helps to memorise bilingual glossaries ⟨4⟩	demo version only limits the number of entries ⟨1⟩	permits to have several glossaries open at the same time ⟨2⟩	-
Final Mark	77	68	67	64	64	60	56	55	27

Table 6: Comparative standalone TMS: *SDL MultiTerm*, *TermX*, *Intragloss*, *AnyLexic*, *Lingo*, *InterpretBank*, *Terminus*, *Intraplex* and *UniLex*.

Feature	**flashTerm** (2015)	**Interpreters' Help** beta (2014)	**ASPLex** (2013)	**WebTerm** 6 (2014)	**Termflow** (2013)	**Acrolinx** (2014)
Manages multiple glossaries ⟨no=0; yes=10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩
N. of possible working languages ⟨≤100=4; >100=7; unlimited=10⟩	>100 ⟨7⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩
N. of languages per glossary allowed ⟨≤3=5; ≥4=10⟩	>4 ⟨10⟩	unlimited ⟨10⟩	6 ⟨10⟩	unlimited ⟨10⟩	>4 ⟨10⟩	unlimited ⟨10⟩
N. of descriptive fields ⟨non=0; 1=3; [2-5]=7; >5=10⟩	>5 ⟨10⟩	unlimited ⟨10⟩	>5 ⟨10⟩	>5 ⟨10⟩	>5 ⟨10⟩	3 ⟨7⟩
Handles documents ⟨no=0; yes=10⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	yes ⟨10⟩	no ⟨0⟩
Remote Glossary Exchange ⟨no=0; yes=10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩
Unicode compatibility ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩
Imports from ⟨1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	-	MS Word, Excel, Open/Libreoffice & CSV ⟨4⟩	MS Word, Excel & other CAT formats ⟨5⟩	> 5 ⟨5⟩	> 5 ⟨5⟩	> 5 ⟨5⟩
Exports to ⟨non=0; 1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	> 5 ⟨5⟩	Excel & PDF ⟨2⟩	MS Word, Excel & other CAT formats ⟨5⟩	> 5 ⟨5⟩	CSV & TBX ⟨2⟩	> 5 ⟨5⟩
Embedded online search for translation candidates ⟨no=0; yes=5⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩
Interface's supported languages ⟨1=1; [2-5]=3; >5=5⟩	English & Deutsch ⟨2⟩	English ⟨1⟩	English+3 ⟨3⟩	> 5 ⟨5⟩	English & Deutsch ⟨2⟩	English ⟨1⟩
Well-documented ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩	yes ⟨5⟩
Availability ⟨proprietary without demo=1; proprietary with demo=3; free=5⟩	proprietary with demo ⟨3⟩	free ⟨5⟩	proprietary with demo ⟨3⟩	proprietary without demo ⟨1⟩	proprietary without demo ⟨1⟩	proprietary with demo⟨3⟩
Other relevant features ⟨subjective analysis=max. 5⟩	clean interface and allows to add illustrations ⟨3⟩	clean and straightforward TMS that allows to add an unlimited number of translation terms ⟨5⟩	-	-	-	simple interface and allows to add illustrations ⟨3⟩
Final Mark	78	77	76	76	75	74

Table 7: Comparative web-based TMS: *flashTerm*, *Interpreters' Help*, *ASPLex*, *WebTerm*, *Termflow* and *Acrolinx*.

And finally, mobile terminology applications, or TMS apps are undoubtedly the next step in this ever-evolving domain of term management. TMS apps are systems which have been developed or optimised for small handheld devices, such as mobile phones, smartphones, iPads or PDAs, among others. Some of the most popular ones are Glossary Assistant[93] and The Interpreter's Wizard[94]. Table 8 provides a comparative summary of the main features that characterise these two mobile TMS. Again, overall punctuations have also been offered to serve as a quick guide or checklist for interpreters.

Feature	**Glossary Assistant** 1.2 (2015)	**The Interpreters' Wizard** 2.0 (2011)
Manages multiple glossaries ⟨no=0; yes=10⟩	yes ⟨10⟩	yes ⟨10⟩
N. of possible working languages ⟨≤100=4; >100=7; unlimited=10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩
N. of languages per glossary allowed ⟨≤3=5; ≥4=10⟩	10 ⟨10⟩	2 ⟨5⟩
N. of descriptive fields ⟨non=0; 1=3; [2-5]=7; >5=10⟩	non ⟨0⟩	non ⟨0⟩
Handles documents ⟨no=0; yes=10⟩	no ⟨0⟩	no ⟨0⟩
Unicode compatibility ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩
Imports from ⟨1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	Plain Text ⟨1⟩	Proprietary Format ⟨1⟩
Exports to ⟨non=0; 1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	Plain Text ⟨1⟩	non ⟨0⟩
Embedded online search for translation candidates ⟨no=0; yes=5⟩	no ⟨0⟩	no ⟨0⟩
Interface's supported languages ⟨1=1; [2-5]=3; >5=5⟩	English ⟨1⟩	English ⟨1⟩
Remote Glossary Exchange ⟨no=0; yes=5⟩	no ⟨0⟩	no ⟨0⟩
Well-documented ⟨no=0; yes=5⟩	yes ⟨5⟩	no ⟨0⟩
Availability ⟨proprietary without demo=1; proprietary with demo=3; free=5⟩	proprietary with demo ⟨3⟩	free ⟨5⟩
Operating System(s) ⟨1=1; 2=3; ≥3=5⟩	Android and Windows ⟨3⟩	iOS (iPad) ⟨1⟩
Other relevant features ⟨subjective analysis=max. 5⟩	user-friendly and intuitive interface ⟨4⟩	quick performance ⟨1⟩
Final Mark	53	39

Table 8: Comparative mobile TMS: *Glossary Assistant* and *The Interpreter's Wizard*.

To sum up, web-based programs obtained higher *average score* when compared with standalone and mobile TMS (76, 60 and 46, respectively). These results can be explained by the companies' effort and the cutting-edge technology used during their development. Another fact that contributes to the increasing interest in web-based TMS is that nowadays companies are more orientated towards developing centralised systems in order to provide uniform services to both staff and clients. Nevertheless, this effort requires higher investment in equipment and manpower to maintain these systems and consequently make them more expensive compared with standalone or mobile TMS. Despite mobile TMS do not get acceptable scores when compared with standalone and web-based TMS and they do not offer the necessary comfort to manage terminology, they still play an important role when a quick search for terminology is required, e.g. while in a booth.

4.2.3 Final Remarks

This section summarised the work reported in Costa et al., 2014a, Costa et al., 2015b and Costa et al., 2017. Various TMS currently available on the market as well as an

[93] `http://swiss32.com`
[94] `http://the-interpreters-wizard.appsios.net/`

overview of the most relevant features that these tools should have in order to help interpreters before and during the interpretation process were presented. In detail, seventeen TMS were presented and compared with the aim of assessing them on the basis of a set of 15 features previously identified and a scoring system. The results obtained can be used to guide interpreters when choosing specific tools for a given interpretation project, i.e. the TMS that would best cater for their specific needs, in order to help them work more efficiently, store and share terminology more easily, as well as save time when looking for a specific feature most suited to a specific interpreting service.

Our main findings suggest that most TMS are not envisaged to be used by interpreters. Therefore, TMS do not fulfil completely the needs of this group of end-users as regards speed of consultation, intuitive navigation, possibility of updating the terminology record in the interpretation booth, freedom to define the basic structure, multiple ways of filtering data and sharing information, etc. Conversely, those tools devoted to interpreters (and mainly developed by interpreters) are fairly basic and only include a limited number of features.

Given that quality terminology management is a top priority for interpreters, there seems to be a pressing need to design TMS tailored specifically to assist interpreters both prior and during their interpreting services. In this vein, it would be necessary to ascertain interpreters' terminology needs (as opposed to translators'), and then, devote more efforts to the development of web-based, standalone and, particularly, mobile TMS in order to provide on-site consultation of glossaries, terminologies, lists of proper names and conversion figures, etc. No doubt, technology-assisted interpreting will offer a challenging and fruitful research niche for many years to come.

4.3 Translators' attitudes towards Terminology Extraction Tools

This section summarises the work reported in the article Costa et al., 2016b, which aimed at investigating translators' attitudes towards Terminology Extraction Tools (TET) and at identifying a set of desirable features that can be used to help translators choosing the most adequate TET for a given translation task. It is important to mention that this work was conducted in collaboration with Anna Zaretskaya, the other EXPERT Early Stage Researcher (ESR) at the University of Malaga and main author of "Translators' requirements for translation technologies: a user survey" (Zaretskaya et al., 2015), which we used as a starting point for this research and extended with the ideas reported in Costa et al., 2014a.

The purpose of TET is to help users build terminological resources in a (semi-)automatic way. The need for such resources comes mostly from the growing needs in information management and translation, which make it more and more necessary to have some automated assistance when performing terminology-related tasks. Companies, freelancers and professionals in various linguistic fields can resort to these tools to, for example build glossaries, thesauri and terminological dictionaries that they use directly in their work. Moreover, terminology extraction (TE) is embedded in a number of Natural Language Processing (NLP) and linguistic research tasks, such as automatic indexing, Machine Translation (MT), Information

Extraction (IE), creation of ontologies and knowledge bases, and corpus analysis. Although they have such broad range of applications, these tools are often designed for one specific purpose, which consequently makes their usage challenging when employed in a different setting. Moreover, not every TET offers a full set of desirable features and settings, which makes it sometimes challenging to find the perfect tool for the task in hand. Apart from the functionalities they offer, TET also differ as to the environment they work in. For instance, standalone installable tools require an installation process and work as independent computer programs. There also exist web-based tools, which work within a browser. And finally, there are reusable software that facilitates the development of larger applications, called frameworks.

Considering the existing variety of TET and corresponding offered set of features, it is not clear how a professional translator is to proceed when choosing a TET suitable for the job. As we will see further, there are various TET that are specifically created for translators. But do they have all the necessary characteristics for translators? What exactly are these characteristics? And, furthermore, how can we choose the most suitable TET for a given task? In the next sections we will report the translators' attitudes towards TET and analyse 9 different tools, in an attempt to answer the aforementioned questions.

4.3.1 Terminology Extraction Tools (TET)

In Costa et al., 2016b we presented various TET divided into three different categories, standalone, web-based and reusable TE libraries, named frameworks. Standalone software is probably the most popular type of software today, and TET are no exception. SDL MultiTerm Extract[95], Simple Extractor[96] and TermSuite[97] are just three examples of standalone TET. The advantages are that web-based TET, compared to standalone tools, do not require any prior installation as they can be accessed within a web browser. Although most of web-based TET are often integrated as features in cutting-edge web-based applications with a wider purpose, such as managing corpora or terminology (e.g. Sketch Engine[98] and Terminus[99], respectively), there also exist tools like the TET by Translated[100], which was developed with the proper purpose of TE. Different from the other two types of tools, frameworks are not complete software products but reusable software environments or libraries that can be used or even completely integrated in larger translation software applications, products or solutions. In particular, systems of this type are often used in Information Retrieval (IR), where identification and indexing of terminology serves as an aid to information retrieval queries. In detail, the purpose of TE for both information retrieval and document retrieval is to isolate terms that contain enough informational content to support retrieval based on the queries supplied when querying a set of documents. Examples of TE frameworks are Keyphrase Extraction Algorithm[101] (Kea), Rainbow[102] and Java Automatic Term

[95] http://www.translationzone.com/products/multiterm-extract/
[96] http://www.dail-software.com/help/9_en/index.html
[97] http://termsuite.github.io
[98] https://www.sketchengine.co.uk
[99] http://terminus.iula.upf.edu
[100] http://labs.translated.net/terminology-extraction/
[101] http://www.nzdl.org/kea/
[102] http://okapiframework.org/wiki/index.php?title=Main_Page

Extraction[103] (JATE).

In the next section, we investigate translators' attitudes towards TET and, compare the aforementioned TET using the most useful feature reported by the Zaretskaya et al., 2015 survey's participants.

4.3.2 Translators' Preferences and Opinions on the Features of TET

Although translation is one of the most important applications of TE, it has not yet become a common part of the professional translation workflow. This was demonstrated by a user survey replied by over 600 translation professionals, which showed that only 25% of the respondents regularly resorted to terminology extraction in their work (Zaretskaya et al., 2015). This can be due to unsatisfying performance of the existing tools, their interface design, or simply to translators' lack of awareness of these tools and of the benefits they can yield.

TET can differ from each other as to various characteristics, such as their interface type (standalone, web-based or reusable libraries), document formats they support, languages they work with, as well as different search options. According to the survey findings reported in Zaretskaya et al., 2015, 27% of the respondents preferred to have a TE feature within their Computer-assisted Translation (CAT) tool instead of a separate TE software. Some translators, however, preferred a web-based application (9%) or installing a standalone tool on their computer (8%). Nevertheless, the majority (56%) reported that they did not have any preference regarding the tool's interface. The fact that translators prefer to have a TET integrated in their CAT tool is related to the general tendency of CAT tools to include more and more different features. Indeed, translators have to deal with a great number of tools that help them automatise different stages of the translation process, so they prefer having one tool with multiple functions rather than having to look for and in many cases pay for several tools.

Table 9 shows the most useful features that a TET should have according to the survey's participants and which of those are presented in the 9 analysed tools (see Costa et al., 2016b for more details about these TET). The most important feature according to the survey's participants was bilingual term extraction. In fact, considering that within a translation workflow, terminology extraction is performed with the final objective to translate the extracted terms, it is more convenient to have the terms extracted in the two languages simultaneously. The second ranked feature was the possibility to compare the context of the term in the source and the target language, which is another type of bilingual analysis suitable for the translation task. The possibility to validate terms or, in other words, choose the terms that should be extracted instead of extracting all terms was ranked third and is also considered useful for translators. Compiling a bilingual dictionary from parallel texts is another useful feature. Finally, the respondents considered it useful to extract context together with terms or to see examples from the corpus. Other features that were considered included support for different file formats, possibility to sort terms by frequency, support for many languages, possibility to specify the minimal number of occurrences of the words, show linguistic information about the

[103] https://github.com/ziqizhang/jate

	SDL Multiterm	*Simple Extractor*	*TermSuit*	*Sketch Engine*	*Translated s.r.l.*	*Terminus*	*Kea*	*Rainbow*	*JATE*
Bilingual extraction	✓		✓	✓					
Source and target context comparison	✓								
Terms validation	✓	✓		✓		✓	✓	✓	✓
Bilingual dictionaries compilation	✓		✓						
Context extraction	✓	✓	✓		✓	✓	✓	✓	✓
Support various file formats	✓	✓	✓	✓		✓	✓	✓	✓
Rank terms by frequency	✓	✓	✓	✓		✓	✓	✓	✓
Support for many languages	✓		✓	✓		✓	✓	✓	✓
Specify the minimal number of occurrences	✓	✓		✓		✓	✓	✓	✓
Show linguistic information	✓		✓			✓			
Specify the maximum number of translations			✓						
Stopword list option	✓	✓			✓		✓	✓	✓
Choose the minimum and maximum number of words per term	✓	✓					✓	✓	✓
Term statistics	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 9: Comparison chart of features.

term, and select the maximum number of translations for one term. All of them were considered useful, but were not among the most useful features. And finally, some features were not considered so important by the respondents. One of them was the stopword list option, i.e. allow to choose whether to use a stopword list, and others use it by default. Choosing the minimum and the maximum number of words per term, which was also among the least useful features, can be tuned by all the mentioned TE frameworks, for example. And finally, term statistics, which to some extent are provided by all tools, were not very important for most translators either.

4.3.3 Final Remarks

Although terminology extraction plays an important role in several disciplines such as linguistic research or language teaching, it is in the field of translation, particularly in the translation industry where its advantages are fully exploited and integrated in the workflow. An example of that is the use of bilingual term extraction, compiling dictionaries and comparing context in different languages as essential features for translators' work. In addition, it is also very useful for translators to see the terms in their context in order to understand their meaning and be able to find an adequate translation equivalent. Not all existing tools, however, provide these functionalities. After a careful analysis of the priorities for the design and features to be included in a TET, we identified 14 main features. Then, based on these features, we made a comparative analysis of 9 TET currently available on the market.

A next step in the right direction could be to gather detailed information to better ascertain translators' technology awareness and then standardise these 14 features into a discriminative scoring system, similarly to what we have done in section 4.2. Although our work already highlighting some of the features that translators can expect from some of the TET currently available on the market, a standardised scoring system would be of great help, not only to guide translators choosing a specific tool for a given project, which would help them work more efficiently, but also help those responsible for improving and/or designing new tools.

4.4 Summary

This chapter summarised the research carried out in Costa et al., 2014b; 2015b; 2016b and Costa et al., 2017. Each section explored the second Research Question (RQ2), discussed in section 1.2 from different perspectives. In detail, this chapter has presented an overview of tools and applications available for interpreting practice and training, as well as for translation practice.

Unlike translators, for whom a myriad of computer-assisted tools are available, interpreters have not benefited from the same level of automation or innovation. As shown in Costa et al., 2014b; 2015b and summarised in section 4.1, their work relies by and large on traditional or manual methods. Although the number of these technologies is growing fast due to an increasing interest towards interpreters' needs, they are still insufficient and unable to fulfil all the necessary requirements. After exploring and cataloguing various technology-assisted interpreting tools available on the market (TMS, note-taking applications for consecutive interpreting, voice recording applications and training tools), we concluded that there is an urgent need to develop technologies that automate the process, increase the productivity and ease the labour-intensive activities of an interpreter before and during an interpretation service.

In an attempt to help interpreters choosing the TMS that best caters for their specific needs, section 4.2 focused on identifying and establishing a set of measurable features that permit them to assess and distinguish the different TMS concerning individual's and/or company's needs in such a way that the results would be useful for both potential customers as well as to the designers of such systems. After a careful analysis of the features that interpreters expect from a TMS and evaluation through our standardised scoring system, we found out that none of the seventeen

evaluated TMS fulfil completely the needs of this group of end-users as regards speed of consultation, intuitive navigation, possibility of updating the terminology record in the interpretation booth, freedom to define the basic structure, multiple ways of filtering data and sharing information, etc. In fact, those tools devoted to interpreters are fairly simple, only include a limited number of features and are not envisaged to be used by interpreters. Accordingly, there is a pressing need to ascertain interpreters' terminology needs (which are different to translators'), and devote more efforts to the deployment of better and more complete TMS.

Finally, section 4.3, focused exclusively on ascertain translators' most favourite terminology extraction features. After a careful analysis of the priorities for the design and features to be included in a TET, we successfully identified fourteen main features. Apart from the functionalities these TET have to offer, we also categorised them according to the environment they work in (i.e. there are standalone installable tools, web-based tools, and reusable software named frameworks). Based on translators' most favourite terminology extraction features, we made a comparative analysis of various well-known TET currently available on the market and concluded that none of them, however, fulfil all the translators' needs, which consequently prevents the vast majority of professional translators to adopt them in their daily work.

To sum up, technology-assisted tools open up a new world of possibilities for both interpreters and translators. Nevertheless, not all existing tools either TMS or TET, however, fulfil all interpreters' and translators' needs. A next step in the right direction could be to gather detailed information to better ascertain their' technology awareness and real needs in order to design new tools or improve existing ones.

Chapter 5

Assessing Comparable Corpora

"I cannot teach anybody anything,
I can only make them think."

—Socrates

This chapter describes the studies reported in Costa et al., 2015a; Costa, 2015; Zampieri et al., 2015; Costa et al., 2015c and Costa et al., 2016a. Each publication explores different aspects of the third Research Question (RQ3) discussed in section 1.2. Apart from these research publications, throughout the chapter various programs and tools created during this work are described in detail (see sections 5.1.2.1 and 5.2.2.3).

Corpus linguistics lacks strategies for describing comparable corpora. Currently most descriptions of comparable corpora are textual and, questions such as "what sort of a comparable corpora is this?", "how does their documents are related with each other?", "should we always trust on (semi-)automatic tools to compile specialised comparable corpora?", or "how to improve the comparability of the corpus?" for example, can only be answered impressionistically. In an attempt to answer the aforementioned questions, this chapter aims at exploring the issue of assessing the internal degree of similarity in comparable corpora of unknown composition by replacing subjectivity with mathematical proofs and accurate measurements.

Section 5.1 describes various Distributional Similarity Measures (DSMs) and Natural Language Processing (NLP) techniques used to build an efficient methodology capable of measuring and ranking documents based on their similarity scores. Moreover, this section also summarises the way the research was conducted in practice through various interconnected studies (Costa, 2015; Costa et al., 2015c; 2016a), as well as the data used in these research experiments. In short, these studies focus on describing and comparing different types of documents and on evaluating how DSMs perform the task of filtering out noisy documents.

Section 5.2 introduces a system capable of computing the similarity between two sentences (Costa et al., 2015a). Similarity measures play a crucial role in various areas of text processing and translation technologies ranging from improving Information Retrieval (IR) rankings and text summarisation to Machine Translation (MT) evaluation and enhancing matches in Translation Memory (TM). Despite computing the semantic similarity between sentences remains a complex and difficult task, we built and submitted for evaluation a system to the annual SemEval'15 task 2: Semantic Textual Similarity.

Section 5.3 presents a system capable of discriminate between similar languages (Zampieri et al., 2015). Although language identification can be considered by some people as a solved task, recent studies have shown that language identification systems often fail to achieve satisfactory performance across different datasets and domains, particularly with datasets containing short pieces of texts such as *tweets*, code-switching data, or when discriminating between very similar languages (e.g. European Portuguese and Brazilian). Given these challenges, we decided to build a system capable of discriminating between similar languages and language varieties. In order to evaluate its performance, we submitted the system to the 2015 Discriminating between Similar Languages (DSL) shared task.

It is important to mention that in the end of each section we report our main research findings and give some cues for further improvements (sections 5.1.5, 5.2.4 and 5.3.4). To complete the chapter, a general discussion and main findings about the results are presented in section 5.4.

5.1 Assessing, Measuring and Ranking Documents in Comparable Corpora

The EAGLES - Expert Advisory Group on Language Engineering Standards Guidelines (EAGLES, 1996b) defined "comparable corpora" as follows: "a comparable corpus is one which selects similar texts in more than one language or variety. There is as yet no agreement on the nature of the similarity, because there are very few examples of comparable corpora". Since 1996, when this definition was given, many comparable corpora have been compiled, analysed and employed in a wide range of disciplines, since it has become an essential resource in several research domains such as Natural Language Processing (NLP), terminology, language teaching, Machine Translation (MT), amongst others (cf. Sharoff, 2007; Corpas Pastor, 2008; Leturia et al., 2009; Gonçalo Oliveira et al., 2010; Snover et al., 2011; Skadiņa et al., 2012; Sharoff, 2013; Tan et al., 2014; Vela and Tan, 2015; Rapp et al., 2016). Therefore, at this point we can state that there are no more "very few examples of comparable corpora". As "comparable corpora are seen as answering perceived needs for texts as examples of 'natural' original text in the source language culture" (Maia, 2003), it is not surprising that in the last decades we have witnessed an increased interest for these resources and a great boost in their usage in NLP applications and by language users.

Nevertheless, as far as we are concerned, "there is as yet no agreement on the nature of the similarity" (EAGLES, 1996b). The uncertainty about the data we are dealing with is still an inherent problem to those who deal with comparable corpora in their research. Indeed, little work has been done on automatically characterising such linguistic resources (Kilgarriff, 2001; Skadiņa et al., 2010a; Sharoff, 2013; Köhler, 2013), and attempting a meaningful description of their content is often a perilous task (Corpas Pastor and Seghiri, 2009). A corpus is usually tagged just with a short description of its content, such as "casual speech transcripts" or "tourism specialised comparable corpus", along with some specifications about authorship, date, source, number of documents, tokens and types. In our view, such tags that usually comes along with the corpus are of little use to those users seeking for a representative and/or high quality domain-specific corpora or to those who are interesting in re-using those corpora for other purposes. As a result, most of the resources at our disposal are built and shared without deep analysis of their content, and those who use them blindly trust on the people's or research group's name behind their compilation process, with lack of real knowledge about the relatedness quality of the corpus or, in other words, how similar the documents are.

Word frequency and co-occurrence lists play a pivotal role when exploring and exploiting comparable corpora. They form a compact summary of what is in a corpus. They also make it possible to assess how similar two sentences and/or documents are, and how they contrast with each other (Costa et al., 2015a; Agirre et al., 2016). Despite Maia in her article in 2013 (Maia, 2003) could not help but conclude that "comparability is in the eye of the beholder", her conclusion is not a satisfactory state of affairs. Moreover, we do not want the sampling for the datasets underlying our scientific endeavour to be subjective (Kilgarriff, 2010). Instead, we want to replace subjectivity with mathematical proofs and accurate measurements.

Accordingly, this section aims at presenting and evaluating an innovative

methodology capable of measuring how comparable, or similar documents within a comparable corpus are. By taking advantage of several statistical methods presented in the literature and by exploiting available NLP technologies (section 5.1.1), we propose a new methodology capable of assessing and measuring how the documents correlate with each other in terms of content (section 5.1.2). Then, in section 5.1.3 we describe the data used through the various interconnected studies, which are summarised in section 5.1.4. Finally, the main findings and ideas for further improvements are presented in section 5.1.5.

5.1.1 Distributional Similarity Measures (DSMs)

Although the task of structuring information from unstructured natural language texts is not an easy task, NLP in general and, Information Retrieval (IR) (Singhal, 2001) and Information Extraction (IE) (Grishman, 1997) in particular have been making some progress in the way the information is accessed, extracted and represented. In detail, IR and IE play a crucial role in the task of locating and extracting specific information within a collection of documents or other natural language resources according to some request. To do so, these two NLP tasks mainly take advantage of a large number of statistical methods based on words and their co-occurrence. Essentially, their methods aim at finding the most frequently used words and treating the rate of usage of each word in a given text as a quantitative attribute. Then, these words serve as features for a given statistical method. Following Harris' distributional hypothesis (Harris, 1970), who assumes that similar words tend to occur in similar contexts, these statistical methods are suitable, for instance to find similar sentences based on the words they contain (Costa et al., 2015a), to automatically extract and validate semantic entities from corpora (Costa et al., 2010; Costa, 2010; Costa et al., 2011), or even to compare two corpora between each other (Kilgarriff, 2001). To this end, it is assumed that the amount of information contained in a document could be evaluated by summing the amount of information contained in the document words. And the amount of information conveyed by a word can be represented by means of the weight assigned to it (Salton and Buckley, 1988; Turney and Pantel, 2010; Baroni, 2013; Baroni et al., 2014). Accordingly, we tested two IR measures commonly used in the literature, the Spearman's Rank Correlation Coefficient (SCC) and the Chi-Square (χ^2) to compute the similarity between documents written in the same language (see section 5.1.4). We chose these two measures because they are independent of text size (mostly because both use a list of the common entities), as well as language-independent (see sections 5.1.1.1 and 5.1.1.2 for more information).

The SCC distributional measure (section 5.1.1.1) has been shown effective on determining similarity between sentences, documents and even on corpora of varying sizes (Kilgarriff, 2001; Costa et al., 2015a). It is particularly useful, for instance to measure the textual similarity between two documents because it is easy to compute and is independent of text size as it can directly compare ranked lists for large and small texts.

The χ^2 similarity measure (section 5.1.1.2) has also shown its robustness and high performance. By way of example, χ^2 have been used to analyse the conversation component of the British National Corpus (Rayson et al., 1997), to compare corpora (Kilgarriff, 2001; Köhler, 2013), and to identify topic related clusters in imperfect

transcribed documents (Ibrahimov et al., 2002). It is a simple statistic measure that permits to assess whether relationships between two variables in a sample are due to chance or, on the contrary, the relationship is systematic.

For all these reasons, DSMs in general and SCC and χ^2 in particular have a wide range of applicabilities (cf. Lee, 1999; Kilgarriff, 2001; 2010; Köhler, 2013; Sharoff, 2013). In this vain, this chapter aims at proving that these simple, yet robust and high-performance text size and language-independent measures allow us to describe the relatedness between documents in comparable corpora (see section 5.1.4). Hereafter, we describe in detail the theoretical and mathematical concepts behind the SCC and χ^2 measures (sections 5.1.1.1 and 5.1.1.2, respectively).

5.1.1.1 Spearman's Rank Correlation Coefficient (SCC)

In this work, the SCC was adopted and calculated as in Kilgarriff (2001). Firstly, a list of the common entities[104] L between two documents d_l and d_m is compiled, where $L_{d_l,d_m} \subseteq (d_l \cap d_m)$. It is possible to use the top n most common entities or all common entities between two documents, where n corresponds to the total number of common entities considered $|L|$, i.e. $\{n|n \in \mathbb{N}^0, n \leq |L|\}$ –in this work we used all the common words for each document pair, i.e. $n = |L|$. Then, for each document the list of common entities (e.g. L_{d_l} and L_{d_m}) is ranked by frequency in an ascending order ($R_{L_{d_l}}$ and $R_{L_{d_m}}$), where the entity with lowest frequency receives the raking position $\mathsf{1}$ and the entity with highest frequency receives the numerical raking position n. In the case of ties in rank, where more than one entity in a document occurs with the same frequency, the average of the ranks is assigned to the tying entities. For instance, if the entities e_a, e_b and e_c had the same frequency and ranked in the $\mathsf{6}^{th}$, $\mathsf{7}^{th}$ and $\mathsf{8}^{th}$ position, all three entities would be assigned the same rank of $\frac{6+7+8}{3} = \mathsf{7}$. Finally, for each common entity $\{e_1, ..., e_n\} \in L$, the difference in the rank orders for the entity in each document is computed, and then normalised as a sum of the square of these differences $\left(\sum_{i=1}^{n} s_i^2\right)$. The final SCC equation is presented in expression 5.1, where $\{SCC|SCC \in \mathbb{R}, -1 \leq SCC \leq 1\}$.

By a way of example let e_x be a common entity (i.e. $\{e_x\} \in L$) and $R_{L_{d_l}} = \{1\#e_{n_{d_l}}, 2\#e_{n-1_{d_l}}, ..., n\#e_{1_{d_l}}\}$ and $R_{L_{d_m}} = \{1\#e_{n_{d_m}}, 2\#e_{n-1_{d_m}}, ..., n\#e_{1_{d_m}}\}$ the resulting ranked list of common words for d_l and d_m, respectively. Assuming that e_x is the $3\#e_{n-2_{d_l}}$ and $1\#e_{n_{d_m}}$, i.e. e_x is in the 3^{rd} position in $R_{L_{d_l}}$ and in the 1^{st} position in $R_{L_{d_m}}$, s would be computed as $s_{e_x}^2 = (3-1)^2$ and the result would be 4. Then, this process is repeated for the remain $n-1$ entities and the resulted SCC score will be seen as the similarity value between d_l and d_m.

$$SCC(d_i, d_j) = 1 - \frac{6 * \sum_{i=1}^{n} s_i^2}{n^3 - n} \tag{5.1}$$

[104] In this work, the term "entity" refers to "single words", which can be a token, a lemma or a stem.

5.1.1.2 Chi-Square (χ^2)

The χ^2 measure also uses a list of common words (L). Similarly to SCC, it is also possible to use the top n most common entities or all common entities between two documents, and again in this work we use all the common words for each document pair, i.e. $n = |L|$. The number of occurrences of a common words in L that would be expected in each document is calculated from the frequency lists. If the size of the document d_l and d_m are N_l and N_m and the entity e_i has the following observed frequencies $O(e_i, d_l)$ and $O(e_i, d_m)$, then the expected values are $e_{i_{d_l}} = \frac{N_l * (O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$ and $e_{i_{d_m}} = \frac{N_m * (O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$. Equation 5.2 presents the χ^2 formula, where O is the observed frequency and E the expected frequency. The resulted χ^2 score should be interpreted as the interdocument distance between two documents. It is also important to mention that $\{\chi^2 | \chi^2 \in \mathbb{R}, 1 \leq \chi^2 < +\infty\}$, which means that the more unrelated the common words in L are, the lower the χ^2 score will be.

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{5.2}$$

Table 10 shows a contingency table example on the assumption that we have two common entities e_i and e_j between two documents d_l and d_m (i.e. $L = \{e_i, e_j\}$).

	d_l	d_m	*Total*
e_i	O=11 E=8.08	O=4 E=6.92	15
e_j	O=3 E=5.92	O=8 E=5.08	11
Total	14	12	26

Table 10: Example of a contingency table.

This table contains: i) the observed frequencies (O); ii) the totals in the margins; iii) and the expected frequencies (E), which are obtained by applying the following formula: $\frac{column\ total}{N} * row\_total$, e.g. $E(e_i, d_l) = \frac{14}{26} * 15 = 8.08$. After writing down the expected frequencies in the table, we are ready to calculate the χ^2 score (see Equation 5.3).

$$\frac{(11 - 8.08)^2}{8.08} + \frac{(3 - 5.92)^2}{5.92} + \frac{(4 - 6.92)^2}{6.92} + \frac{(8 - 5.08)^2}{5.08} = 5.41 \tag{5.3}$$

In order to find out the p value associated with the obtained $\chi^2 = 5.41$, we need to: a) calculate the degrees of freedom (df), where $df = \frac{(\#rows-1)}{(\#columns-1)} = \frac{(2-1)}{(2-1)} = 1$; b) and, search for the p value in the abbreviated table of Critical Values for the χ^2 test. In this example, the obtained value of $\chi^2 = 5.41$ with $df = 1$ exceeds the cut-off of 3.84 shown on the table at the 0.05 level. Therefore, $p < 0.05$, which allow us to reject the *null* hypothesis. Thus, the result means that $(1, N = 26) = 5.41, p < 0.05$.

5.1.2 Methodology

This section aims at presenting a simple, yet efficient methodology capable of measuring and ranking documents based on their similarity scores. This methodology was firstly presented in Costa, 2015 and improved afterwards in Costa et al., 2015c and Costa et al., 2016a. As we will see in the following sections, this methodology will allow us not only to measure and rank documents, but also to describe and extract information about the corpus in hand and the degree of relatedness in it (section 5.1.4). Moreover, part of the methodology's pipeline was also successfully used in the task of computing the semantic similarity between two sentences (see section 5.2). Hereafter, we describe the entire pipeline along with all the tools, libraries and frameworks used in the process. It is important to mention that the entire methodology's pipeline was successfully deployed in two Java programs, which were made publicly available and, thus free for being used by anyone, both in a research or in a commercial setting (see section 5.1.2.1 for more details).

i) **Data Preprocessing**: firstly, the data is processed with the OpenNLP[105] Sentence Detector and Tokeniser. Then, the annotation process is carried out with the TT4J[106] library, which is a Java wrapper around the popular TreeTagger (Schmid, 1995), a tool specifically designed to annotate text with Part-of-Speech (POS) and lemma information. Regarding the stemming, we use the Porter stemmer algorithm provided by the Snowball[107] library. A method to remove punctuation and special characters within the words was also implemented. Finally, in order to get rid of the noise, a stopword list[108] was compiled to filter out the most frequent words in the corpus. Once a document is computed and the sentences are tokenised, lemmatised and stemmed, our system creates a new output file with all this new information, i.e. a new document containing the original, the tokenised, the lemmatised and the stemmed text. Using the stopword list mentioned above a Boolean vector describing if the entity is a stopword or not is also added to the document. This way, the system is able to use only the tokens, lemmas and stems that are not stopwords.

ii) **Identifying the list of common entities between documents**: in order to identify a list of Common Entities (hereafter, CE), a co-occurrence matrix is built for each pair of documents. Only those that have at least one occurrence in both documents are considered. As required by the DSMs (see section 5.1.1), their frequency in both documents is also stored within this matrix ($L_{d_l,d_m} = \{e_i, (f(e_i, d_l), f(e_i, d_m)); e_j, (f(e_j, d_l), f(e_j, d_m)); ...; e_n, (f(e_n, d_l), f(e_n, d_m))\}$, where f represents the frequency of an entity in a given document d). With the purpose of analysing and comparing the performance of different DSMs, three different lists are created to be used as input features: the first one using the Number of Common Tokens (NCT),

[105] https://opennlp.apache.org
[106] http://reckart.github.io/tt4j/
[107] http://snowball.tartarus.org
[108] Freely available to download through the following URL: https://github.com/hpcosta/stopwords.

the second one using the Number of Common Lemmas (NCL) and the third one using the Number of Common Stems (NCS).

iii) **Computing the similarity between documents**: the similarity between documents is calculated by applying three different DSMs ($DSMs = \{DSM_{CE}, DSM_{SCC}, DSM_{\chi^2}\}$, where CE, SCC and χ^2 refer to the number of Common Entities, Spearman's Rank Correlation Coefficient and Chi-Square, respectively), each one calculated by using three different input features (NCT, NCL and NCS).

iv) **Computing the document final score**: the document final score $DSM(d_l)$ is the mean of the similarity scores of the document with all the documents in the collection, i.e. $DSM(d_l) = \frac{\sum_{i=1}^{n-1} DSM_i(d_l, d_i)}{n-1}$, where n refers to the total number of documents in the collection and $DSM_i(d_l, d_i)$ the resulted similarity score between the document d_l with all the documents in the collection.

v) **Ranking documents**: finally, the documents are ranked in a descending order according to their DSMs scores (i.e. DSM_{CE}, DSM_{SCC} or DSM_{χ^2}).

5.1.2.1 Deployed Software (DSMModule and PreProcessor)

As mentioned in the introductory section 1.3 of this thesis, apart from the research publications, various programs and tools were created during this work. Since the beginning, one of the proposed objectives was to contribute for the advancement of science in general and computational linguistics in particular, either through scientific publications or deployed software tools to help language users automatise their daily tasks. To do so, we implemented the entire methodology's pipeline and made publicly available the resulted Java code to be used by anyone, both in a research or in a commercial setting.

Accordingly, we successfully implemented the entire methodology's pipeline in a Java program called DSMModule[109] (Distributional Similarity Measures Module). In short, the DSMModule is an open source program that aims at offering the user with a simple, yet efficient set of algorithms capable of measuring and ranking either sentences or documents based on their similarity scores. It implements all the tools, libraries and frameworks mentioned in the previous section 5.1.2 in a single program with the purpose of accessing, measuring and ranking documents based on their shared content, and consequently help researchers decide whether a specific document should be integrated in the corpus or not.

For those only interested in processing and annotating raw textual data, we also thought about them and specifically deployed a simpler version of the DSMModule, named PreProcessor[110]. Despite various POS taggers, Lemmatisers, Stemmers, Named Entities Recognisers, Sentence Splitters, Tokenisers and Stopword Checkers can be used for this purpose, they are independent programs built for a specific purpose (e.g. identify the word's stem). Thus, when users want to use more than one or import them in their own programs/applications, their integration turns to be really complex and time-consuming. As an attempt to help the user solving this

[109] `https://github.com/hpcosta/DSMModule`
[110] `https://github.com/hpcosta/PreProcessor`

problem, we created the PreProcessor. It offers the user with a simple, yet robust and agile variety of morphosyntatic options to process and annotate raw textual data.

5.1.3 The INTELITERM Comparable Corpus

Regarding that most of our experiments used a specific comparable corpus, we decided to dedicate one section to present and describe it, so it would be easier to understand its content, locate its information and avoid duplicate information through the next section 5.1.4.

The INTELITERM[111] corpus is a domain comparable corpus composed of documents retrieved from the Internet –it is important to mention that although INTELITERM also has a small parallel subcorpus, it was not considered for the task in hand. It was firstly compiled manually by researchers with the purpose of building a representative noise-free domain-specific corpus for the Spanish, English, German and Italian Tourism and Beauty domain. Nevertheless, in order to boost the size of this domain corpus, automatic compilation was also necessary and, at a second stage, more documents were automatically retrieved with the BootCaT[112] compilation tool (Baroni and Bernardini, 2004). In both compilations the same variables and external criteria were followed in order to maintain the homogeneity and the quality of the full corpus (see Table 11).

Criterion	**Description**
Temporal	The date of publication or creation of the texts selected is as recent as possible.
Geographical	All the texts selected are geographically limited, that is, all the English, German Italian and Spanish texts used are from UK, Germany, Italy and Spain respectivly, so as to avoid possible diatopical terminological variation, such as the Spanish spoken in Mexico or Venezuela.
Formal	The texts selected pertain to a specialised communicative setting, that is, a medium-high level of specialisation, are originally written in the languages of the study and are in their full electronic format.
Genre or textual typology	All the texts selected belong to the the same genre, that is, promotional tourism texts retrieved from the Internet, containing products and services wellness and beauty.
Authorship	All the texts are authentic documents drafted by relevant authors, institutions or companies.

Table 11: Variables and external criteria used during the compilation process.

Regarding the INTELITERM comparable corpus structure, it can be divided in four subcorpora according to the working languages of the project: English, Spanish, German and Italian. Within these subcorpora it can be further divided by type of document, that is: manually collected original texts, manually collected translations and automatically collected original texts. Accounting for the purpose of this work, the entire corpus is analysed, i.e. all the original and translated documents manually compiled for English (*i_en_od* and *i_en_td*), Spanish (*i_es_od* and *i_es_td*), German

[111] http://www.lexytrad.en/

[112] `http://bootcat.sslmit.unibo.it`

(*i_de_od* and *i_de_td*) and Italian (*i_it_od* –the researchers did not find translated documents for Italian), as well as the documents automatically compiled by the researchers with the *b*oot*c*aT compilation tool for English, Spanish, German and Italian (*bc_en*, *bc_es*, *bc_de* and *bc_it*, respectively). All the information about these subcorpora is presented in Table 12. In detail, this table shows the number of documents (nD), the number of types (types), the number of tokens (tokens), the ratio of types per tokens ($\frac{types}{tokens}$) per subcorpus and, its source type (Source Type), which can be original, translations or crawled from the Internet (original, translation and crawled, respectively). These values were obtained by using the corpus analysis toolkit for concordancing and text analysis software Antconc 3.4.3 (Anthony, 2014).

	nD	types	tokens	$\frac{types}{tokens}$	Source Type
i_en_od	151	11,6k	496,2k	0.023	original
i_en_td	60	6,9k	83,1k	0.083	translation
i_es_od	224	13,0k	207,3k	0.063	original
i_es_td	27	3,4k	16,4k	0.207	translation
i_de_od	138	21,4k	199,8k	0.049	original
i_de_td	109	5,5k	26,8k	0.205	translation
i_it_od	150	19,9k	386,2k	0.051	original
bc_en	111	41,1k	563,5k	0.073	crawled
bc_es	246	32,8k	735,4k	0.045	crawled
bc_de	253	58,3k	482,4k	0.121	crawled
bc_it	122	11,9k	81,5k	0.147	crawled

Table 12: Statistical information about the various INTELITERM subcorpora.

5.1.4 Experiments

After presenting our methodology, the existing gap that needed to be explored and the data in hand, it is time to join the pieces in a test scenario and explain our findings. For this purpose, we summarised in the next four sections the work reported in Costa et al., 2015c; Costa, 2015 and Costa et al., 2016a. We used the methodology proposed in section 5.1.2 and the various DSMs described in section 5.1.1 to assess the INTELITERM corpus presented in section 5.1.3. Firstly, the various subcorpora that were manually compiled (containing original and translated texts) were explored and the content of original documents with the translated ones was compared in order to understand how it differ from each other from a statistical point of view (section 5.1.4.1). Then, we present an experiment where we explored how the translated documents affect the general relatedness scores when merged with the original ones (section 5.1.4.2). To do so, we randomly selected and added different percentages of translated documents to the original subcorpora. Similarly to these two previous experiments, sections 5.1.4.3 and 5.1.4.4 report how the documents (semi-)automatically compiled relate with those manually compiled and how the average scores vary when adding different amounts of documents (semi-)automatically crawled from the Web to the various subcorpora manually compiled, respectively. Finally, in section 5.1.4.5 we summarise the work we did on evaluating how DSMs perform the task of filtering out noisy documents, i.e.

documents with a low level of relatedness. To do so, we injected different sets of out-of-domain documents, randomly selected from a different corpus, and evaluated the DSMs' accuracy.

In order to perform these experiments over the INTELITERM corpus, we applied the methodology explained in section 5.1.2 and three different DSMs, i.e.: the number of Common Entities (CE); the Spearman's Rank Correlation Coefficient (SCC); and, the Chi-Square (χ^2). As an input features to these DSMs, three different types of entities obtained from the corpus were used (i.e. tokens, lemmas and stems). Figures 11, 12 and 13 show the Number of Common Tokens (NCT) between documents on average (av), the SCC and the χ^2 scores along with the associated standard deviations (σ - vertical lines extending from the bars) per measure and subcorpus (i.e. original, translated and automatically compiled with BootCaT).

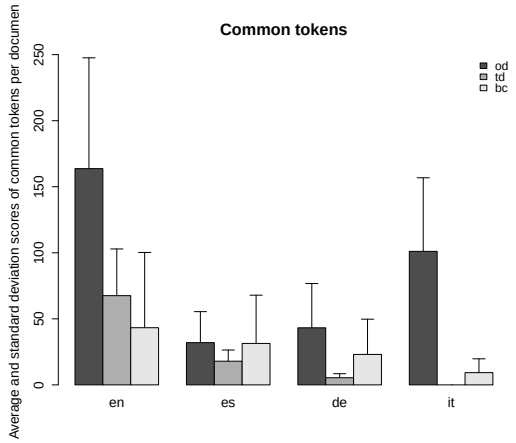


Figure 11: Common tokens average and standard deviation per subcorpora.

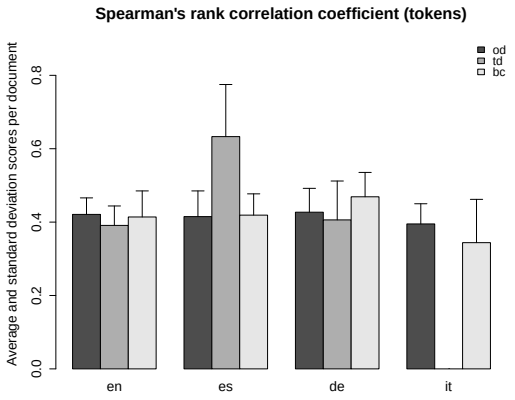


Figure 12: SCC average and standard deviation scores per subcorpora.

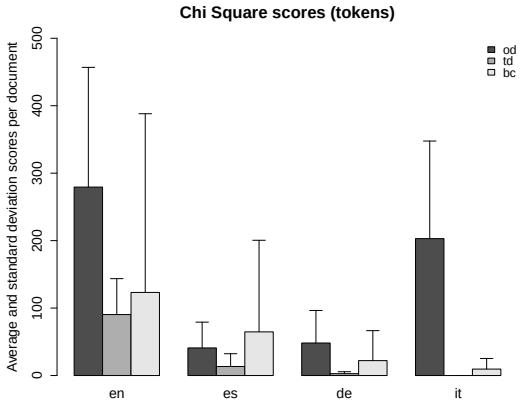


Figure 13: χ^2 average and standard deviation scores per subcorpora.

At this point, it is worth mentioning that for the following experiments the whole corpus in its original size and form was employed (not just a sample) and, therefore, all the obtained results and findings come from the entire population, that is the INTELITERM: English (*i_en_od*, *i_en_td*, *bc_en*), Spanish (*i_es_od*, *i_es_td*

and *bc_es*), German (*i_de_od*, *i_de_td* and *bc_de*) and Italian (*i_it_od* and *bc_it*) subcorpora. –Please note that the Italian subcorpus does not have translated documents.

5.1.4.1 How Original and Translated Documents Relate between each other

This section summarises part of the work reported in Costa et al., 2015c; Costa, 2015 and Costa et al., 2016a. More precisely, this section is dedicated to the analysis of the various INTELITERM manually compiled subcorpora, i.e. original versus translated documents. Figures 14 to 22 present the resulted average scores per document in a box plot format for all the combinations DSM versus feature for the various INTELITERM subcorpora. Each box plot displays the full range of variation (from min to max), the likely range of variation (the interquartile range), the median and the high maximums and low minimums (also known as outliers).

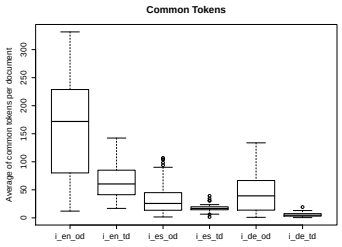


Figure 14: Feature: Tokens. Average of NCT per document (original vs. translated).

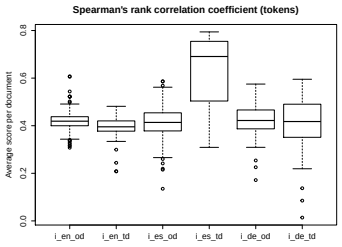


Figure 15: Feature: Tokens. SCC average scores per document (original vs. translated).

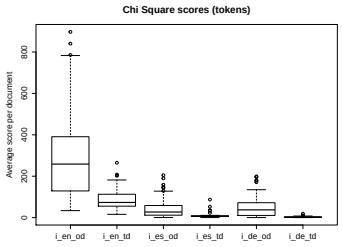


Figure 16: Feature: Tokens. χ^2 average scores per document (original vs. translated).

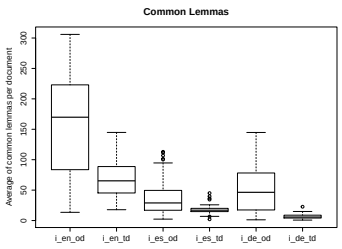


Figure 17: Feature: Lemmas. Average number of Lemmas per document (original vs. translated).

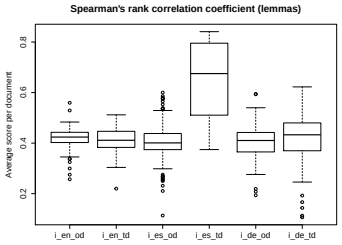


Figure 18: Feature: Lemmas. SCC average scores per document (original vs. translated).

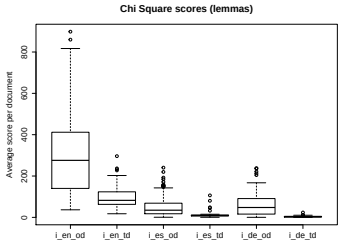


Figure 19: Feature: Lemmas. χ^2 average scores per document (original vs. translated).

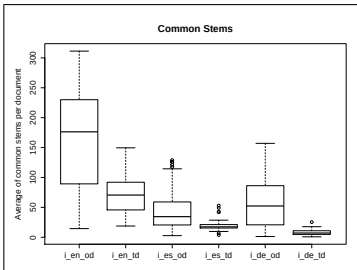


Figure 20: Feature: Stems. Average number of Stems per document (original vs. translated).

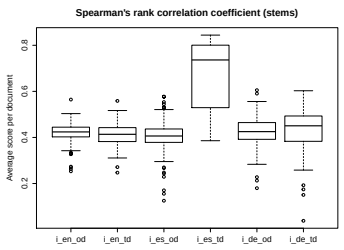


Figure 21: Feature: Stems. SCC average scores per document (original vs. translated).

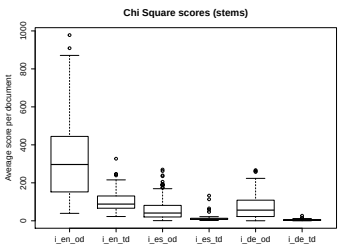


Figure 22: Feature: Stems. χ^2 average scores per document (original vs. translated).

The first observation we made from our data was that the distributions between the features resulted quite similar (see Figures 14, 17 and 20). And, although it is not possible to generalise these results to other types of corpora or domains, all the DSMs suggested the same finding for the data at stake: it is possible to achieve acceptable results only by using raw words (i.e. tokens). As stems and lemmas require more processing power and time to be used as features –especially lemmas due to the Part-of-Speech (POS) tagger dependency and time consuming process implied, the possibility of using only tokens results to be an advantage.

Based on the achieved results (see Figures 14, 15 and 16), we stated that the scores for each subcorpus are symmetric (roughly the same on each side when cut down in the middle), which means that the data in hand is normally distributed. There were some exceptions, such as the SCC and χ^2 average scores for the *i_es_td* and for the *i_de_td*, which will be discussed later on in this section. Another interesting observation is related to the high number of Common Entities (CE) –see Figures 11, 14, 17 and 20– in the original documents (*i_en_od*, *i_es_od* and *i_de_od*) when compared with translated documents (*i_en_td*, *i_es_td* and *i_de_td*, respectively). Our assumption to this phenomenon was based on the fact that these documents were collections of translated documents (translated from different languages and by different translators) retrieved from the Internet, and, consequently, due the variability of several linguistic features, such as vocabulary, style, repetition, language sources, etc. found in each text, the number of CE between the documents was much lower.

Although the Number of Common Tokens (NCT) per document on average is higher for the *i_en_od* subcorpus, the interquartile range (IQR) is larger than for the other subcorpora (see Figures 11 and 14), which means that the middle 50% of the data is more distributed and thus the average of NCT per document is more variable. Moreover, longest whiskers (the lines extending vertically from the box) in Figure 14 could indicate variability outside the upper and lower quartile. Therefore, we could state that the *i_en_od* subcorpus had a big variety in the types of documents and consequently some of them were only roughly correlated to the rest of the subcorpora. Nevertheless, the data is skewed left, which means that the majority is strongly similar, i.e. the documents have a high degree of relatedness between each other. This idea was sustained by the positive average SCC scores presented in Figure 15 and the set of outliers founded above the upper whisker. Moreover, the

average of 0.42 SCC score and σ=0.045 also implies a strong correlation between the documents in the *i_en_od* subcorpus. Regarding the χ^2 scores, the longest whisker outside the upper quartile in Figure 16 also indicates an high degree of relatedness between the documents. Regarding the *i_en_td* subcorpus, the NCT, the SCC and the χ^2 scores (Figures 14, 15 and 16) and the average of 67.54 common tokens per document and σ=35.35 (Figure 11) suggested that the data is normally distributed (Figure 15) and the documents are –not as much as the *i_en_od* subcorpus, yet– also highly related between each other.

Among all the subcorpora, the *i_es_od* subcorpus is the biggest one with 224 documents (Table 12). Nevertheless, Figures 11 and 14 reveal a lower NCT compared with both English subcorpora. Although a deep linguistic analysis would give us a more accurate explanation, a theoretical approach for this phenomenon is that Spanish has a richer morphology compared to English. Therefore, due to a higher number of inflected forms per lemma, there is a larger number of tokens and consequently less common tokens per document in Spanish. When analysing Figures 14 and 16, the box plots for the *i_es_od* subcorpus look similar to the *i_en_td* when shifted up. Except for the longest whisker observed in Figure 15, the SCC scores also show similar distributions, averages and standard deviations when compared with the *i_en_td* subcorpus (see Figure 11).

Despite the German *i_de_od* subcorpus has more types and less tokens (21.4k and 199.8k, respectively) when compared with the *i_es_od* (13k types and 207.3k tokens), their $\frac{types}{tokens}$ ratio does not vary much from each other (0.049 against 0.063, see Table 12 for more details) as well as the NCT, SCC and χ^2 scores. For example, the NCT between the documents on average for the *i_es_od* subcorpus is 31.97 and the σ=23.48, against an av=43.21 and a σ=33.52 for the *i_de_od* subcorpus. Furthermore, their SCC and χ^2 average and standard deviation scores are even more expressive (i.e. *i_es_od*'s SCC={av=0.415; σ=0.07} vs. *i_de_od*'s SCC={av=0.427; σ=0.065} and *i_es_od*'s χ^2={av=40.922; σ=38.212} vs. *i_de_od*'s χ^2={av=48.235; σ=45.301}).

As observed in Figures 14, 15 and 16, the average scores per document for both *i_es_td* and *i_de_td* subcorpora are slightly different from the *i_en_td* box plots. Apart from the low NCT per document, the χ^2 standard deviations is higher than their averages (*i_es_td*={av=13.40; σ=18.95} and *i_de_td*={av=2.771; σ=2.883}), and from the expressive *i_es_td*'s SCC variability inside and outside the IQR indicates some inconsistency in the data. This instability has been explained by the low number of types (*i_es_td*=3.4k and *i_ed_td*=5.5k) and tokens (*i_es_td*=16.4k and *i_de_td*=26.8k) and their 0.207 and 0.205 $\frac{types}{tokens}$ ratio (Table 12). As mentioned by Baker, 2006, the $\frac{types}{tokens}$ ratio tends to be useful when looking at relatively small documents, and in this specific case these subcorpora only have on average 607 and 246 tokens (*i_es_td*=$\frac{16400}{27}$ $\approx$ 607 and *i_de_td*=$\frac{26800}{109}$ $\approx$ 246), and 126 and 50 types per document (*i_es_td*=$\frac{3400}{27}$ $\approx$ 126 and *i_de_td*=$\frac{5500}{109}$ $\approx$ 50), which made them an excellent test case. When compared with the low ratios from the other subcorpora (see Table 12), –even for this type of corpora– these ratios can be considered high. In this context, a high ratio suggests that a more diverse form of language is employed, which can also explain the low NCT and χ^2 scores for these subcorpora. By contrast, a low ratio can indicate a great number of repetitions (the same word occurring again and again), likely indicating a relatively narrow range of subjects. Despite the high SCC, the data is asymmetric and variable (large IQR) –see Figure 15. This

happened because most of the common entities had a low frequency in the documents and consequently they ranked close together in the ranking lists, which resulted in high SCC scores mostly because of the resulted high value in the numerator (see Equation 5.1).

To sum up, this first experiment was dedicated to the analysis of the various INTELITERM manually compiled subcorpora (original versus translated) and the main findings were:

i) the DSMs input features provided similar scores;

ii) the original documents resulted to have a higher number of common entities when compared with the translated ones;

iii) and, the DSMs suggested that the English and the Italian original subcorpora were composed by documents with a higher degree of relatedness in comparison with the rest of the subcorpora.

The next section reports how the translated documents affect the general relatedness degree when merged with the original subcorpora.

5.1.4.2 How Translated Documents affect the General Relatedness Degree when Merged with the Original Documents

This section summarises part of the work reported in the article Costa et al., 2016a. After analysing the various original and translated INTELITERM subcorpora, the next obvious step was to understand how the translated documents would affect the general relatedness scores when merged with the original ones. To do so, we performed an experiment in which we randomly selected and added different percentages of translated documents to the original subcorpora. More precisely, we added 10%, 20%, 30% and 100%[113] to the various original subcorpora. Figures 23, 24 and 25 show the resulted average scores per document for each percentage. As expected, the more documents are injected, the lower the NCT is (see Figure 23). Apart from that, there are a couple of interesting findings that come out from this exercise. Although the NCT for Spanish decreased when the entire set of translated documents (100%) was mixed with the original subcorpus ≈9.3% less common tokens per documents, the drop was not significant. In fact, the average scores per document increased ≈1.19% and ≈1.22% when 20% and 30% of the translated documents were added, respectively, in relation to the original subcorpus. The SCC and χ^2 scores also corroborate this fact (Figure 24 and 25, respectively) as they do not vary much when different set of documents were added. We observed a similar phenomenon for English, where the original subcorpus has an av=163.70 and when 10%, 20%, 30% and 100% of the translated documents were added, the NCT only decreased ≈3.2%, ≈3.4%, ≈6.1% and ≈23.6%, respectively. In particular for those cases where a bigger specialised subcorpus is required to conduct research, we concluded that –even if it means having some noisy documents within the collection– based on the statistical findings, the original and translated Spanish and/or English subcorpora could be merged together without highly compromising their original subcorpora's general relatedness degrees –especially for Spanish where the general relatedness score only droped ≈9.3% when added 27 translation documents (which corresponds to an increase of ≈12% of documents). Despite the NCT decreased

[113]The number of documents that correspond to these percentages can be inferred from Table 12.

≈23.6% for English when added 60 translation documents, the increase was bigger than for Spanish, more precisely by ≈39.7%.

Among all the subcorpora, the German was the one that seem to maintain a bigger gap between its subcorpora. In other words, when both subcorpora were merged together, the general relatedness score drastically decreased by ≈53.4%. These facts are pretty visible in Figures 23 and 25.

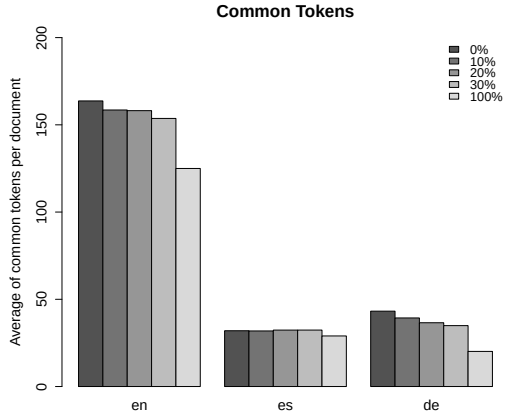


Figure 23: Average NCT per document after adding translated documents to the original subcorpora.

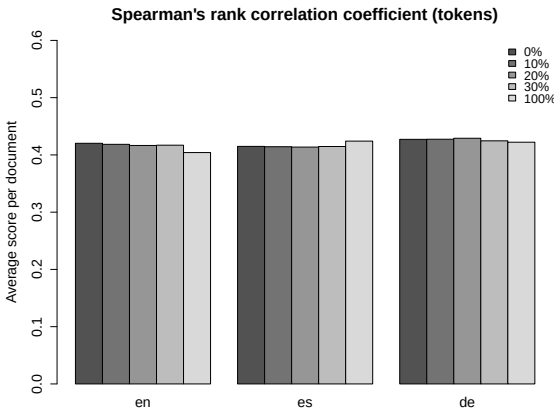


Figure 24: SCC average scores per document after adding translated documents to the original subcorpora.

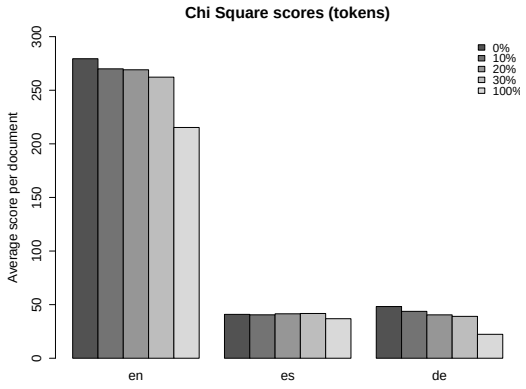


Figure 25: χ^2 average scores per document after adding translated documents to the original subcorpora.

In the end, we concluded that if a bigger specialised subcorpus would be required for Spanish and/or English, the theoretical and statistical evidences showed that both original and translated subcorpora could be merged without dramatically decrease their internal similarity relatedness degree –especially for Spanish the drop would be just ≈9.3%. Nevertheless, we stated that it is always advisable to perform research on the original subcorpus and only if a bigger subcorpus is required for the task in hand, proceed with the merge of the corresponding translated subcorpus.

Similarly to the previous section, the next section is dedicated to the comparison between the documents manually compiled with those (semi-)automatically compiled.

5.1.4.3 How (Semi-)automatic and Manually Compiled Documents Relate between each other

This section summarises part of the work reported in the article Costa et al., 2016a. Similarly to section 5.1.4.1, this section reports an experiment dedicated to the comparison between the documents manually compiled with those (semi-)automatically compiled with BootCaT (see section 5.1.3 for more information about the INTELITERM corpus and its subcorpora). Similarly to what we did in section 5.1.4.1, we also performed a statistical comparison between both types of documents in order to understand how their average scores differ from each other.

Figures 26, 27 and 28 put side-by-side the resulted average scores per document in a box plot format for all the working languages (English, Spanish, German and Italian). The first observation we made from Figure 26 was the astonishing different in the NCT between the original and the (semi-)automatic subcorpora for English and Italian. By a way of example, the NCT on average per document for the *i_en_od* subcorpus is 163.70 with a σ=83.89, still, the *bc_en* only has an av=43.28 with a σ=56.97, i.e. ≈74% less common tokens per document on average. In fact, the difference between the Italian subcorpora is even higher, ≈91% less common tokens per document on average to be more precise (*i_it_od*={av=101.08; σ=55.71} and *bc_it*={av=9.26; σ=10.46}). These findings can be corroborated by the higher SCC variation scores and the lower χ^2 scores for both the *bc_en* and the *bc_it* when compared with the *i_en_od* and the *i_it_od* subcorpora, respectively (Figures 27 and 28). Nevertheless, it is worthy to notice that the *bc_en* has a few outliers above the upper whisker, which means that these documents have similar degrees of relatedness to those in the *i_en_od* subcorpus, and thus they should be carefully analysed by the person in charge of the corpus.

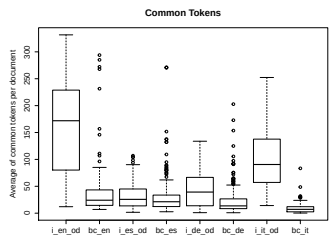


Figure 26: Original vs. (semi-)automatically compiled subcorpora: average NCT per document.

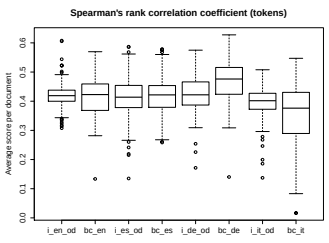


Figure 27: Original vs. (semi-)automatically compiled subcorpora: SCC average scores per document.

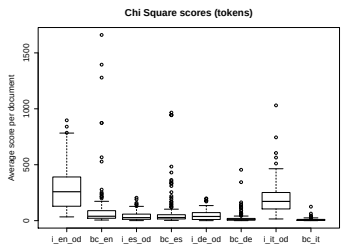


Figure 28: Original vs. (semi-)automatically compiled subcorpora: χ^2 average scores per document.

Regarding the *bc_de* subcorpus, it has ≈22% less common tokens per document on average when compared with the *i_de_od* (*i_de_od*={av=43.21; σ=33.52} and *bc_de*={av=23.06; σ=26.68}). Despite this 22% difference between the German subcorpora, we should not reject by now the hypothesis that these two subcorpora

could not be merged together without dramatically decreasing the internal relatedness degree –nevertheless, a deeper analysis is required as we will see later on in this section. For what concerns the Spanish subcorpora, they look like assembling a similar degree of relatedness between their documents, afterwards their averages and standard deviations do not differ much from each other (*i_es_od*={av=31.97; σ=23.48} and *bc_es*={av=31.38; σ=36.51}). Moreover, their similar SCC and χ^2 scores also seem to point at the same direction (Figures 27 and 28).

In short, on the one hand the DSMs average scores presented in Figures 26, 27 and 28 provide a clear evidence that both manual and (semi-)automatic English and Italian subcorpora do not have much in common. On the other hand, the DSMs average scores suggest that the German and, especially, the Spanish subcorpora could have a similar degree of relatedness between their manually and (semi-)automatically compiled documents, and thus be considered for merge if necessary.

In the next section 5.1.4.4, we put to the test these findings by randomly selecting and adding different percentages of documents from the (semi-)automatically compiled subcorpora to those manually compiled. Our theory was that the general relatedness scores would drop when the documents (semi-)automatically compiled are added. Based on the previous results, a dramatic drop for English and Italian and a smaller decrease for German and particularly for Spanish are expected.

5.1.4.4 How (Semi-)automatic Compiled Documents affect the General Relatedness Degree when Merged with the Original Documents

This section summarises part of the work reported in the article Costa et al., 2016a. After analysing the various original and (semi-)automatic INTELITERM subcorpora, the next obvious step was to understand how the various (semi-)automatic documents would affect the general relatedness scores when merged with the original ones. Thus, we performed an experiment where we randomly selected and added different percentages of (semi-)automatic documents to the original subcorpora.

Figures 29, 30 and 31 show the average scores per document when adding different amounts of documents (semi-)automatically crawled from the Web to the various original subcorpora manually compiled. More precisely, in order to observe how the general relativeness score varies, we randomly selected and increasingly added sets of 10% to the original subcorpora until both subcorpora are completely merged together. Above all, what is important to observe from Figures 29, 30 and 31 is the following: the initial average scores, i.e. the scores of the manually compiled subcorpora (0%); how these scores vary when more documents are added (from 10% to 100%); and the initial and the final scores when both subcorpora are finally merged together (0% and 100%). When Figures 26, 27 and 28 put side-by-side the resulted average scores per document, we already had a clue about what would happen when merging manually with (semi-)automatically retrieved documents, and, in fact, Figures 29 and 31 corroborate the initial thesis. As we can see from Figure 29, the more sets are added, the lower the NCT for all the working languages is.

As mentioned before, the NCT on average per document for the *i_en_od* subcorpus is 163.70. Nevertheless, when the *bc_en* is merged –which means an

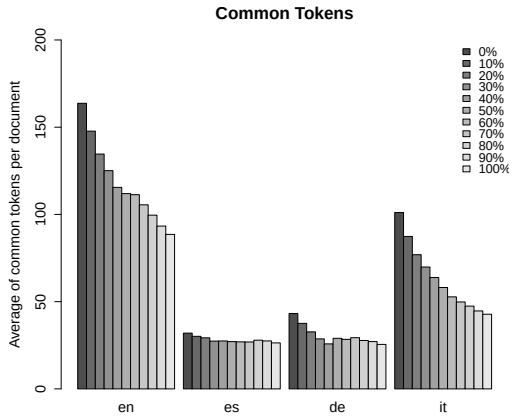


Figure 29: Average NCT per document after adding (semi-)automatic compiled documents to the original subcorpora.

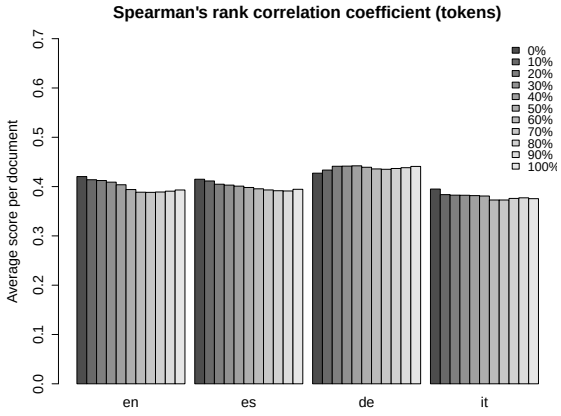


Figure 30: SCC average scores per document after adding (semi-)automatic compiled documents to the original subcorpora.

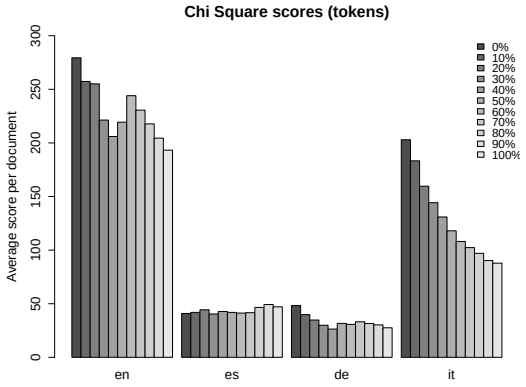


Figure 31: χ^2 average scores per document after adding (semi-)automatic compiled documents to the original subcorpora.

increase of ≈73.5% of documents– the *i_en_od*'s NCT is reduced by almost half (i.e. ≈46% - {*i_en_od* + *bc_en*}={av=88.55}). For Italian the reduction in the NCT is even higher, more precisely ≈58% ({*i_it_od* + *bc_it*}={av=42.79} and ≈81.3% increase in the number of documents). And, German follows this trend with a reduction of ≈41% in the NCT, nevertheless this merge represents an increase of ≈183.3% in the number of documents. The χ^2 scores also point out in the same direction, i.e. the χ^2 scores decrease ≈31%, ≈57% and ≈43% for the {*i_en_od* + *bc_en*}, the {*i_it_od* + *bc_it*} and for the {*i_de_od* + *bc_de*}, respectively. We can observe a similar phenomenon for Spanish in Figure 26. Yet, although the NCT decreases by ≈17% for Spanish when the number of documents increased by ≈103.8%, the degree of relatedness look like somehow stabilises as soon as the first set of documents is added, which means that the *bc_es* subcorpus follows a normal

distribution in terms of content –in this case NCT. Regarding its χ^2, Spanish gets an increase of ≈15%.

Likewise to what we advised in section 5.1.4.2 when we analysed the original versus translated subcorpora, if a bigger specialised subcorpus would be required, in this case, for Spanish –not just only because of the NCT and χ^2 scores from Figures 29 and 31, but also from the previous findings in the previous section 5.1.4.3 (Figures 26, 27 and 28)– the merge of their subcorpora could be performed without dramatically compromise their internal similarity relatedness degree. Or, at least, it would be more advisable than merging the Italian, the German or even the English subcorpora. Although in general the SCC scores drop for three out of the four working languages, they do not, however, are explicit enough to allow us to draw a solid conclusion about them (Figure 30).

In the next section 5.1.4.5, we put to the test both our methodology and the various DSMs in a scenario where different sets of out-of-domain documents are randomly selected and injected from a different corpus. Our theory is that the DSMs would perform well on filtering out these noisy documents.

5.1.4.5 Using DSMs to Filter out Noisy Documents in Comparable Corpora

This section summarises part of the work reported in Costa et al., 2015c. In this experiment we focused on evaluating how DSMs perform the task of filtering out noisy documents, i.e. documents with a low level of relatedness. To do so, we injected different sets of out-of-domain documents, randomly selected from the Europarl[114] corpus[115] (Koehn, 2005) to the original INTELITERM subcorpora. More precisely, we injected 5%, 10%, 15% and 20% to the English, Spanish and Italian original subcorpora (*i_en_od*, *i_es_od* and *i_it_od*, respectively). The number of documents (nD) that correspond to 20% is reported in Table 13 along with their corresponding number of types (types), number of tokens (tokens), and ratio of types per tokens ($\frac{types}{tokens}$) per subcorpus. These noisy documents were randomly selected from the "one per day" Europarl v.7 for three working languages: English, Spanish and Italian (*eur_en*, *eur_es*, *eur_it*, respectively).

	nDocs	types	tokens	$\frac{types}{tokens}$
eur_en	30	3.4k	29,8k	0.116
eur_es	44	5,6k	43,5k	0.129
eur_it	30	4,7k	29,6k	0.159

Table 13: Europarl's statistical information per subcorpus.

After applying the methodology described in section 5.1.2 to these "new twelve subcorpora" (*int_en05*, *int_en10*, ..., *int_it15* and *int_it20*, see Figure 32), we got the documents ranked in a descending order according to their DSMs final scores. As we can see in Figure 32, the more noisy documents are injected, the lower the NCT is. Then, in order to evaluate the DSMs precision, we analysed the first n positions in the ranking lists produced by the three DSMs (individually),

[114] `http://www.statmt.org/europarl/`

[115] Europarl is a parallel corpus composed by proceedings of the European Parliament.

which in this case n corresponds to the number of original documents in a given INTELITERM subcorpus, i.e. *i_en_od*, *i_es_od* and *i_it_od*. Table 14 presents the precision values obtained by the DSMs when injecting different amounts of noise to the various original subcorpora.

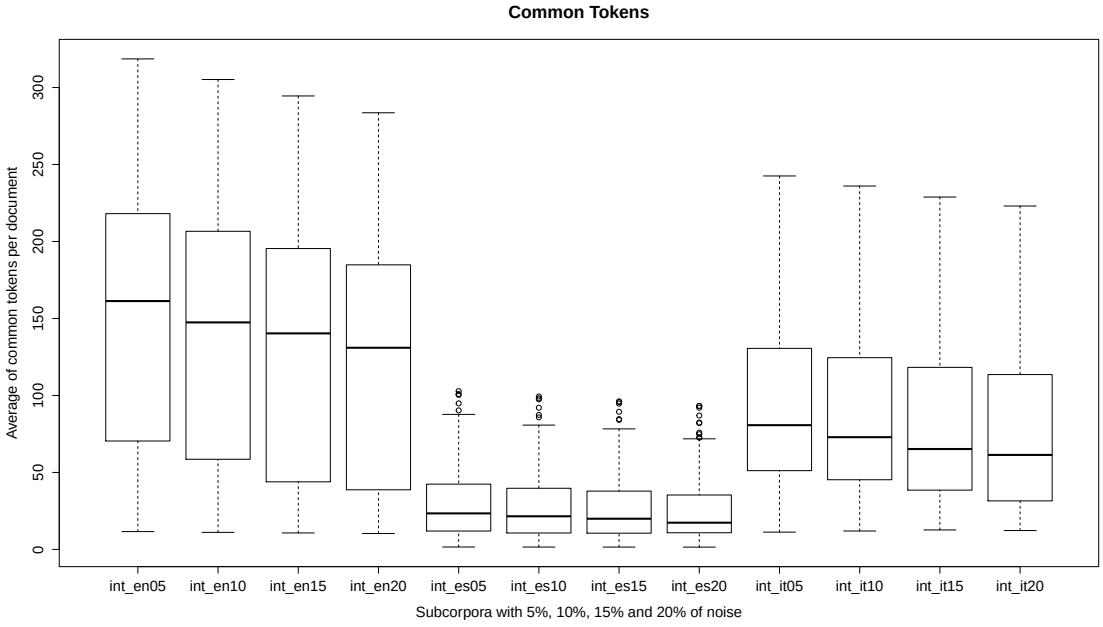


Figure 32: Average scores between documents when injecting 5%, 10%, 15% and 20% of noise to the various INTELITERM subcorpora.

SubC	Noise	NCT	SCC	χ^2
i_en_od	*5%*	0.89	0.22	1.00
	10%	0.73	0.33	1.00
	15%	0.73	0.36	0.95
	20%	0.80	0.37	0.90
i_es_od	*5%*	0.00	0.00	0.38
	10%	0.07	0.07	0.20
	15%	0.09	0.09	0.17
	20%	0.14	0.18	0.23
i_it_od	*5%*	0.88	0.13	0.88
	10%	0.82	0.06	0.82
	15%	0.74	0.09	0.83
	20%	0.73	0.13	0.87

Table 14: DSMs precision when injecting different amounts of noise to the various INTELITERM subcorpora.

As expected, none of the DSMs got acceptable results for Spanish, being incapable of correctly identify noisy documents. However, we need to be aware that this happened due to the pre-existing low level of relatedness between the original documents in the *i_es_od* subcorpus (see section 5.1.4.1 and Costa et al., 2015c for more details). On the other hand, the DSMs showed promising results for English and Italian. By a way of example, the χ^2 reached 100% when we injected

5% and 10% of noise to the int_en subcorpus, and even 90% when injected 20%. Although the NCT got lower precision, in general, when compared with the χ^2, it still reached 80% and 73% when injected 20% of noise to the English and to the Italian subcopora, respectively.

From the evidences shown in Table 14, we concluded that the NCT and the χ^2 were suitable for the task of filtering out low related documents with a high precision degree. The same could not be said about the SCC measure, at least for this specific task.

5.1.5 Final Remarks

This section started by introducing some theoretical concepts about Distributional Similarity Measures (DSMs) and how they can be used to compute the similarity between documents (section 5.1.1). Then, section 5.1.2 described an efficient methodology capable of automatically assess and measure the internal degree of relatedness in comparable corpora. In order to avoid duplicate information, section 5.1.3 was exclusively created to present and describe the data used in our experiments. Next, section 5.1.4 summarised the work reported in Costa et al., 2015c; Costa, 2015 and Costa et al., 2016a into five research experiments. Each experiment addressed the RQ3 (section 1.3) from different perspectives. In detail, the first experiment (section 5.1.4.1) was dedicated to the analysis of the various INTELITERM manually compiled subcorpora (original versus translated). In the end we concluded from a statistical and a theoretical viewpoint that: the various DSMs input features (i.e. tokens, lemmas and stems) provided similar scores; the original documents resulted to have a higher number of common entities when compared with the translated ones; and, the DSMs suggested that the English and the Italian original subcorpora are composed by documents with a higher degree of relatedness in comparison with the rest of the subcorpora. In order to put to the test these findings, the second experiment (section 5.1.4.2) explored how the translated documents affect the general relatedness scores when merged with the original ones. To do so, we randomly selected and added different percentages of translated documents to the original subcorpora. In the end, we concluded that if a bigger specialised subcorpus would be required for Spanish and/or English, the evidences showed that both original and translated subcorpora could be merged without dramatically decrease their internal similarity relatedness degree, especially for Spanish the drop would be smooth. Nevertheless, we advised that it is wise to perform research on the original subcorpus and only if a bigger subcorpus is required, proceed with the merge of the corresponding translated subcorpus. Instead of comparing the original with the translated documents, the third and fourth experiments summarised in sections 5.1.4.3 and 5.1.4.4 compared the original subcorpora with those (semi-)automatically compiled. In detail, section 5.1.4.3 analyses how they relate between each other and section 5.1.4.4 reported how the average scores vary when adding different amounts of documents (semi-)automatically crawled from the Web to the various subcorpora manually compiled. In short, on the one hand the DSMs average scores provided clear evidences that both manual and (semi-)automatic English and Italian subcorpora do not have much in common. On the other hand, the DSMs average scores suggested that the German and, especially the Spanish subcorpora could have a

similar degree of relatedness between the two different type of documents, and thus be considered for merge if necessary. Likewise to what we advised before, if a bigger specialised subcorpus would be required, in this case, for Spanish, the merge of their subcorpora could be performed without dramatically compromise their internal similarity relatedness degree. Or, at least, it would be more advisable than merging the Italian, the German or even the English subcorpora. Finally, the fifth experiment, described in section 5.1.4.5, summarised the work performed on evaluating how the DSMs performed the task of filtering out noisy documents, in this case out-of-domain documents randomly selected from a different corpus. To do so, we injected different sets of out-of-domain documents, randomly selected from the Europarl corpus to the original INTELITERM subcorpora. In the end, none of the DSMs got acceptable results for Spanish, being incapable of correctly identify noisy documents, which we already expected due to the pre-existing low level of relatedness between the original documents. Nevertheless, the DSMs showed promising results for English and Italian. By a way of example, the χ^2 reached 100% when we injected 5% and 10% of noise to the int_en subcorpus, and even 90% when injected 20%. Although the NCT got lower precision, in general, when compared with the χ^2, it still reached 80% and 73% when injected 20% of noise to the English and to the Italian subcopora, respectively. In the end, we concluded that the NCT and the χ^2 could be considered suitable for the task of filtering out low related documents with a high precision degree. The same could not be said about the SCC measure, at least for this specific task.

5.2 Measuring the Semantic Textual Similarity between Sentences

This section summarises part of the work reported in Costa et al., 2015a, which describes the system submitted by the MiniExperts team to the SemEval'15 task 2: Semantic Textual Similarity[116] (STS). It is important to mention that this work was conducted in collaboration with various researchers from the University of Wolverhampton in the UK. Each team element contributed with various ideas, code and features to improve the submitted system. In the end, we made publicly available part of the code used to solve this task.

The task participants were asked to developed a system capable of measuring the STS between two sentences. The idea was to compute how similar two sentences were by returning a similarity score using a scale from 0 (no relation) to 5 (semantic equivalence), and an optional confidence score. In this vain, the MiniExperts team created a system based on a number of linguistically motivated features. It performed satisfactorily for English and obtained a mean 0.7216 Pearson correlation, which ranked 33^{th} among 74. However, it performed less adequately for Spanish, obtaining only a mean 0.5158, which ranked 9^{th} out of 17.

Hereafter, we explain in more detail the goal of the SemEval'15 task 2: STS, our approach and deployed software, the obtained results and, the final remarks and some cues for further improvements (sections 5.2.1, 5.2.2, 5.2.3 and 5.2.4, respectively).

[116] http://alt.qcri.org/semeval2015/task2/

5.2.1 The STS Task

Semantic Textual Similarity (STS) is the task of assigning a real number score to quantify the semantic likeness of two text snippets. Similarity measures play a crucial role in various areas of text processing and translation technologies ranging from improving Information Retrieval (IR) rankings (Lin and Hovy, 2003; Corley and Mihalcea, 2005) and text summarisation to Machine Translation (MT) evaluation and enhancing matches in Translation Memory (TM) and terminologies (Resnik, 1999; Ma et al., 2011; Banchs et al., 2015; Vela and Tan, 2015). However, computing the semantic similarity between sentences remains a complex and difficult task.

The annual SemEval STS task (Agirre et al., 2012; 2013; 2014; 2015; 2016) provides an excellent opportunity for researchers interested in evaluating and comparing their systems' performance on computing how similar two sentences are, using a platform where systems are evaluated on the same data and evaluation criteria. In detail, this SemEval task involved computing how similar two sentences are in both English (Subtask 2a) and Spanish (Subtask 2b). In 2015 the participants were challenged with new datasets in English and Spanish. The English subtask dataset comprised pairs of sentences from news headlines (HDL), image descriptions (Images), answer pairs from a tutorial dialogue system (Answers-student), answer pairs from Q&A websites (Answers-forum), and pairs from a committed belief dataset (Belief). For the Spanish subtask, additional pairs from news and Wikipedia articles were selected. The annotations for both tasks leveraged crowdsourcing.

5.2.2 Approach

Given that each team was allowed to submit three different runs for each task, i.e. English (Subtask 2a) and Spanish (Subtask 2b), we decided to take this opportunity to test and compare different approaches. To do so, we used an improved and revised version of the system submitted to the SemEval'14 (Gupta et al., 2014). As in Gupta et al., 2014, we employed a Machine Learning (ML) method which exploits available NLP technology, adding features inspired by deep semantics (such as parsing and paraphrasing) with Distributional Similarity Measures (DSMs), Conceptual Similarity Measures, Semantic Similarity Measures and Corpus Pattern Analysis[117] (CPA). We used a Support Vector Machine (SVM) in order to compute semantic relatedness for both subtasks. In detail, we built a regression model which estimates a continuous score between 0 and 5 for each sentence pair. The values of C and γ have been optimised through a grid-search which uses a 5-fold cross-validation method, and all systems use an RBF kernel. The system for Subtask 2a (English) was trained on a combination of training and trial data provided by the 2012, 2013 and 2014 SemEval tasks. We used these datasets to form a training set of 9750 sentence pairs combining the different domains covered by the STS task. However, the training set for Subtask 2b (Spanish) was much smaller, at only 804 sentence pairs collected by combining previous datasets.

Hereafter, we briefly describe our approach, i.e. the required preprocessing steps (section 5.2.2.1) and all the features used by our system (section 5.2.2.2).

5.2.2.1 Data Preprocessing

Next, we present all the tools, libraries and frameworks used to preprocess not only the test datasets but also the training datasets.

POS-Tagger, Lemmatiser, Stemmer: the software used for these specific NLP tasks were: the Stanford CoreNLP[118] (Toutanova et al., 2003) toolkit, which provides a lemmatiser, POS-Tagger, Named Entity Recogniser (NER), parsing, and coreference; the TT4J[119] library, which is a Java wrapper around the popular TreeTagger (Schmid, 1995); and the Porter stemmer algorithm provided by the Snowball[120] library.

NER: we took advantage of the Apache OpenNLP library[121] to identify named entities in English and Spanish.

Translation Model: since one of the features we implemented was available only for English (i.e. the Semantic Similarity Measures, see section 5.2.2.2), we trained a Statistical Machine Translation (SMT) system to translate our Spanish dataset into English. To do so, we used the PB-SMT system Moses (Koehn et al., 2007), 5-gram language models with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002),

[117] http://pdev.org.uk
[118] http://nlp.stanford.edu/software/corenlp.shtml
[119] https://code.google.com/p/tt4j
[120] http://snowball.tartarus.org
[121] http://opennlp.apache.org

the GIZA++ implementation of IBM word alignment model 4 (Och and Ney, 2003), with refinement and phrase-extraction heuristics as described in Koehn et al., 2003. We trained this system on the Europarl Corpus (Koehn, 2005) and used Minimum Error Rate Training (Och, 2003) for tuning on the development set.

Resources: given that a number of our features depends on stopwords, we compiled two lists of stopwords[122], one for English and another one for Spanish. We also used two lists (English and Spanish) of candidates for Multiword Expressions (MWEs) as a resource for one of the features. These lists were extracted from the Europarl Corpus (Koehn, 2005) using the collocation modules of the NLTK package (Loper and Bird, 2002), and sorted by the degree of likelihood association between their components (see section 5.2.2.2 for more details).

5.2.2.2 Extracted Features

In addition to the baseline features used in Gupta et al., 2014, we introduced a set of Distributional, Semantic and Conceptual Similarity Measures, as well as a feature reflecting MWEs across sentences. Hereafter, we explain these features in detail.

Baseline Features: the system was built on the baseline system developed for the SemEval'14, which consists of 13 features explained in detail in Gupta et al., 2014. The code which implements these features can be found on GitHub[123].

DSMs: we used two text- and language-independent IR measures, the SCC and the χ^2 to compute the similarity between two sentences written in the same language (see section 5.1.1 for more details about these measures). For every pair of sentences (either English or Spanish), we used the lemmas to extract the list of common terms to compute both measures.

Conceptual Similarity Measures: this feature aims to find the conceptual similarity between two sentences written in the same language. In order to calculate the conceptual similarity, we took advantage of the BabelNet[124] (Navigli and Paolo Ponzetto, 2012) multilingual semantic network. As BabelNet organises lexical information in a semantic conceptual way, we created a conceptual sentence for all input pair of sentences (English and Spanish). More precisely, for every pair of sentence we only extracted lemmatised nouns, verbs, adjectives and adverbs. Then, a conceptual term list was built by extracting all the occurrences of the term in the conceptual network (i.e. BabelNet). As a result, we got a "conceptual representation" of both sentences, each of them containing a set of conceptual term lists. Next, for every term in the "*conceptual_sentence_1*", we counted the number of co-occurrences in the conceptual term lists in the "*conceptual_sentence_2*". In other words, we intersected the terms in *sentence_1* with all the conceptual term lists in *sentence_2*. After computing all the co-occurrences, we used these values to

[122]Both lists are freely available to download through the following URL: `https://github.com/hpcosta/stopwords`.
[123]`https://github.com/rohitguptacs/wlvsimilarity`
[124]`http://babelnet.org`

calculate the Jaccard' (Jaccard, 1901), Lin' (Lin, 1998) and PMI' (Turney, 2001) scores.

Semantic Similarity Measures: this feature takes advantage of the Align, Disambiguate and Walk (ADW)[125] library (Pilehvar et al., 2013), a WordNet-based approach for measuring semantic similarity of arbitrary pairs of lexical items. It is important to mention that this feature is the only one that only works for English, which explains why we have a translation model. In other words, when we are dealing with Spanish text, we use the trained model to translate from Spanish to English (see section 5.2.2.1). As the ADW library permits us to measure the semantic similarity between two raw English sentences, either by using disambiguation or not, we used both options to calculate all the comparison methods made available by the library (WeightedOverlap, Cosine, Jaccard, KLDivergence and JensenShannon divergence).

MWEs: we focused on two more common types of MWEs in English and Spanish, `verb noun` combinations and `verb particle` constructions. Whenever a `verb+noun` or a `verb+particle` combination occurs in our sentence pair, we looked at a prepared list of MWEs, sorted according to their likelihood measures of association, which served as a feature to our system.

5.2.2.3 Deployed Software (STSModule)

As mentioned in the introductory section 1.3 of this thesis, apart from the research publications, various programs and tools were created during this work. As mentioned before, the system was built on the baseline system developed for the SemEval'14 (Gupta et al., 2014), which can be found on GitHub[126]. As soon as we submitted our runs to the SemEval'15 task 2: Semantic Textual Similarity, we decided to upload our this year's additional features of the system on GitHub and made it publicly available so it could be used by anyone. We named our system STSModule[127] (Semantic Textual Similarity Module). In short, the STSModule aims at offering the user with a simple, yet very efficient approach to compute semantic similarity by combining various semantic resources with statistical methods.

5.2.3 Results

The task required the submission of 3 different runs for each task (see Tables 15 and 16).
The runs for the Subtask 2a (English) were identical except for some parameter differences for the SVM training. Our system performed adequately, with our primary run achieving a mean Pearson Correlation of 0.7216. However, the runs for Subtask 2b (Spanish) were trained on different training sets. *Run-1* and *Run-2* were trained on the 804 Spanish sentence-pairs. The Spanish set's *Run-3*, however is trained on the much larger English training set. For this purpose, we needed to translate the Spanish test set into English in order to use the Semantic Similarity

[125] `http://lcl.uniroma1.it/adw`
[126] `https://github.com/rohitguptacs/wlvsimilarity`
[127] `https://github.com/hpcosta/STSModule`

	Run-1	Run-2	Run-3
answers-forums	0.6781	0.6454	0.6179
answers-students	0.7304	0.7093	0.6977
belief	0.6294	0.5165	0.3236
headlines	0.6912	0.6084	0.5775
images	0.8109	0.7999	0.7954
mean	***0.7216***	***0.6746***	***0.6353***
rank (out of 74)	***33***	***45***	***55***

Table 15: Task 2a - Pearson Correlation for English.

	Run-1	Run-2	Run-3
wikipedia	0.5239	0.4671	0.4402
newswire	0.5076	0.5437	0.5524
mean	***0.5158***	***0.5054***	***0.4963***
rank (out of 17)	***9***	***10***	***11***

Table 16: Task 2b - Pearson Correlation for Spanish.

language-dependent features (see sections 5.2.2.1 and 5.2.2.2). This system did not outperform the basic Spanish model used in *Run-1* and *Run-2*, despite the much larger training set. Our Spanish system did not yield a satisfactory performance, achieving a Pearson Correlation score of only 0.5158. This could be part due to the smaller training set in Spanish, and the imperfect translations into English which consequently influenced the performance of the language-dependent features. The detailed results for both tasks are given in the Tables 15 and 16.

5.2.4 Final Remarks

This section presented the work submitted to the SemEval'15 task 2: Semantic Textual Similarity. The MiniExperts team submitted an efficient approach to calculate semantic relatedness for both English and Spanish sentence pairs. We used the same feature set for both tasks, even though it meant translating the Spanish sentences into English before extracting one of the features (i.e. the Semantic Similarity). The system did not performed well for Spanish as it ranked 9^{th} (out of 17), with a 0.5158 average Person correlation over two test sets (0.1747 correlation points less than the best submitted run). On the other hand, it performed reasonably well for English, where the system's best result ranked 33^{th} among 74 submitted runs with 0.7216 Pearson correlation over five test sets (only 0.0799 correlation points less than the best submitted run).

In the future we plan to extract the conceptual description provided by the BabelNet multilingual semantic network in order to match it with the conceptual terms. We have not done that before because we need to treat these descriptions as sentences, which requires filtering out the noise produced by them. Moreover, in order to improve the system's performance, especially for Spanish, it is imperative to increase the sentence-pairs training dataset.

5.3 Discriminating between Similar Languages and Language Varieties

This section summarises part of the work reported in Zampieri et al., 2015, which describes the system submitted by the MMS[128] team to the Discriminating between Similar Languages (DSL) shared task[129] 2015. It is important to mention that this work was conducted in collaboration with researchers from other universities and institutes in Germany (Saarland University and Max Planck Computing and Data Facility). The MMS team participated in the closed submission track using only the dataset provided by the shared task organisers, which contained short texts from 13 similar languages and language varieties. We submitted three runs using different systems and our best system achieved 95.24% accuracy for test set A (containing original texts) and 92.78% accuracy for test set B (containing texts without named entities), which ranked 2^{nd} (out of 9 teams) and 4^{th} (out of 7 teams), respectively. Hereafter, we explain in more detail the goal of the DSL task, our approach, the obtained results and, the final remarks and future plans for improving the submitted system (sections 5.3.1, 5.3.2, 5.3.3 and 5.3.4, respectively).

5.3.1 The DSL Task

Although language identification can be considered by some people as a solved task, recent studies have shown that language identification systems often fail to achieve satisfactory performance across different datasets and domains (Lui and Baldwin, 2011), particularly with: datasets containing short pieces of texts such as *tweets* (Zubiaga et al., 2014); code-switching data (Solorio et al., 2014); or when discriminating between very similar languages (Zampieri et al., 2014).

Given these challenges, the DSL shared task provides an excellent opportunity for researchers interested in evaluating and comparing their systems' performance on discriminating between similar languages and language varieties using short text excerpts extracted from journalistic texts. In detail, the shared task organisers provide all participants with an updated version of the DSL corpus collection v.2.0 (DSLCC) (Tan et al., 2014), a corpus composed of 14 classes, 13 languages[130] and one class containing documents written in previously "unseen" languages to emulate a real-world language identification scenario.

Table 17 presents the languages included in the DSLCC v.2.0 corpus grouped by similarity. The corpus contains 308,000 short text excerpts sampled from journalistic texts (22,000 per class) varying between 20 and 100 tokens per excerpt. It is important to mention that these 22,000 texts per class are divided into 3 partitions, i.e. 18,000, 2,000 and 2,000 instances for training, development and testing, respectively. The test set is further subdivided into two test sets (A and B), each one containing 1,000 instances. While the test set A contains original texts, the organisers replaced named entities for place holders in the set B in order to decrease thematic bias in the classification process.

[128]MMS is an acronym for the team affiliations/locations (Malaga, Munich and Saarland).
[129]`http://ttg.uni-saarland.de/lt4vardial2015/dsl.html`
[130]For the sake of simplicity, we refer to both languages and language varieties as languages.

Language/Variety	Code
Bosnian	*bs*
Croatian	*hr*
Serbian	*sr*
Indonesian	*id*
Malay	*my*
Czech	*cz*
Slovak	*sk*
Brazilian Portuguese	*pt-BR*
European Portuguese	*pt-PT*
Argentine Spanish	*es-AR*
Castilian Spanish	*es-ES*
Macedonian	*bg*
Bulgarian	*mk*
Unknown	*xx*

Table 17: DSL corpus by language and variety.

5.3.2 Approach

Given that each team was allowed to submit three runs, we decided to take this opportunity to test and compare different approaches. To do so, we developed three different systems based on team MMS-member's previous work in language identification and related tasks. The first two systems were previously used for the Native Language Identification (NLI) (Gebre et al., 2013) and the third one has been applied to language variety identification. Hereafter we briefly describe the three systems and the their corresponding submission runs:

- ◇ **Run-1 - Logistic Regression with TF-IDF[131] Weighting**: we opted for the Logistic Regression using the LIBLINEAR open source library (Fan et al., 2008) from scikit-learn (Pedregosa et al., 2011) and fix the regularisation parameter to 100.0. This regression algorithm has proved its efficiency before in different classification problems including for example temporal text classification (Niculae et al., 2014).

- ◇ **Run-2 - SVM with TF-IDF Weighting**: we used a Support Vector Machine classifier (Joachims, 1998). This approach delivered a slightly better performance than Logistic Regression during the NLI shared task. On a very challenging dataset containing TOEFL essays written by speakers of 11 different languages, TF-IDF with SVM reached 81.4% and 84.6% accuracy on the test set when using 10-fold cross validation.

- ◇ **Run-3 - Likelihood Estimation**: we used a simple, yet efficient and fast method that combines Laplace smoothing and a probabilistic classifier. This approach was previously applied to distinguish Brazilian and European Portuguese texts (Zampieri and Gebre, 2012) and it is available as an open source tool called *VarClass* (Zampieri and Gebre, 2014).

[131]TF-IDF is an acronym for Term Frequency - Inverse Document Frequency.

5.3.3 Results

Table 18 reports the official shared task results in terms of accuracy and highlights the best results for each dataset. As we can see, the results obtained by the three systems are very similar. Nevertheless, the SVM with TF-IDF Weighting approach obtained slightly better overall performance (*Run-2*). As we expected, the systems' performance drops from test set A to test set B. This means that our systems rely on named entities to discriminate between similar languages. It is important to point out that we did not do any specific training with the blinded named entities. Probably we could have achieved better results if we had prepared our systems to cope with this variation.

Run	**Test Set A**	**Test Set B**
Run-1	94.09%	**92.77%**
Run-2	**95.24%**	**92.77%**
Run-3	94.07%	92.47%
Rank	2^{nd} *out of 9*	4^{th} *out of 7*

Table 18: Official shared task overall accuracy results.

Table 19 presents the accuracy obtained by our best system (SVM with TF-IDF Weighting - *Run-2*) for each of the 14 classes. The results show that our system achieved perfect performance in two of the language groups (Czech/Slovak and Bulgarian/Macedonian), probably due to exclusive characters present in one of the languages, as well as in identifying the "unseen" languages in test set A. Although the performance did not drop for Croatian and Malay when comparing test set A and B as it did for the rest of the languages, we do not think that this reflects any property of Croatian nor Malay nor any characteristics of the dataset. This is a simple preference of the classifier when distinguishing Croatian from Bosnian and Serbian, and Malay from Indonesian.

Language/Variety	**Test Set A**	**Test Set B**
Bosnian	83.5%	76.6%
Croatian	91.8%	92.2%
Serbian	93.9%	90.7%
Indonesian	99.2%	97.5%
Malay	99.4%	99.5%
Czech	100%	99.9%
Slovak	100%	100%
Brazilian Portuguese	93.6%	90.5%
European Portuguese	93.0%	86.7%
Argentine Spanish	91.2%	89.2%
Castilian Spanish	94.8%	94.5%
Macedonian	100%	100%
Bulgarian	100%	100%
Unknown	100%	99.8%

Table 19: *Run-2* (SVM with TF-IDF Weighting): performance per language.

5.3.4 Final Remarks

This section presented the work submitted to the Discriminating between Similar Languages (DSL) shared task. We submitted three different approaches to deal with the task of discriminating between similar languages and language varieties, and their overall scores turned out to be very similar. The linear SVM classifier combined with TF-IDF Weighting (*Run-2*) achieved slightly better results than the other two methods, i.e. 95.24% against 94.07% and 94.09% accuracy on test set A. The system ranked 2^{nd} (out of 9 teams) on the test set A and 4^{th} (out of 7 teams) on the test set B.

The systems' performance drop from test set A to test set B, which was already expected because named entities play an important role in this kind of task. One of the ways to cope with the influence of named entities in text classification is to use delexicalised text representations relying on POS tags or hybrid representations mixing word forms and grammatical categories. In previous works, however, the results obtained using POS tags to discriminate between Spanish varieties, indicate that the use of more abstract text representations do not result in performance gain (Zampieri et al., 2013).

In the future we would like to return to the question of text representation and investigate whether we can propose features that deliver higher performance across multiple datasets. An interesting approach would be to model these three systems hierarchically. This would result in a two-level classification task, first identifying the language group (grouped by similarity) and then the language itself. This approach has been already proposed by the NRC team, the DSL winner of the 2014 edition (Goutte et al., 2014), and thus, we think this idea is worthy of further investigation.

5.4 Summary

This chapter summarised various research experiments reported in Costa et al., 2015a; Costa, 2015; Zampieri et al., 2015; Costa et al., 2015c and Costa et al., 2016a. Each experiment explored the third Research Question (RQ3), discussed in section 1.2, from a different perspective. Apart from the experiments, this chapter also introduces the various programs and tools created throughout this work (see sections 5.1.2.1 and 5.2.2.3 for more details).

As suggested by Köhler, 2013, the notions of "comparison" and consequently that of "comparability" are predicates with at least three arguments: comparable (A, B, C), where A represents the object to be compared, B the object A is compared with, and C is the respect with which A and B are compared to each other. Further arguments might represent the purpose of the comparison and consequently affect the criteria of comparison. This simple analysis is important to understand the aim of this chapter and the central importance that the DSMs play in the comparability formula. A given document may be comparable to another one with respect to the frequency distribution of letters but not with respect to its genre or its length, it may be comparable with respect to its publication data but not with respect to its topics. In this work, we rather focused on assessing the degree of comparability in comparable documents and sentences according to the content they share between each other.

The first part of the chapter, section 5.1, focused on presenting the theoretical background about Distributional Similarity Measures (DSMs) and Natural Language Processing (NLP) techniques used to build a methodology capable of measuring and ranking documents based on their similarity scores. Through various experiments we demonstrated that this methodology can be used not only to measure and rank documents, but also to describe and extract information about comparable corpora and the degree of relatedness of its documents (see section 5.1.4). Moreover, we also evaluated how the DSMs performed the task of filtering out noisy documents, in this case out-of-domain documents randomly selected from a different corpus. Despite the SCC resulted incapable of filtering out noisy documents, it played an important role describing the data in hand. On the other hand, the NCT and the χ^2 demonstrated to be efficient in both tasks.

Then, sections 5.2 and 5.3 reported two research systems, built in collaboration with researcher from other Universities, which were submitted for evaluation to the SemEval'15 task 2: Semantic Textual Similarity and to the Discriminating between Similar Languages (DSL) shared task 2015. Regarding the first task on measuring the semantic textual similarity between two sentences, we used part of the methodology described in section 5.1.2. Apart from that the DSMs, we exploited other measures, such as Conceptual Similarity Measures and Semantic Similarity Measures in order to train a regression model to compute the semantic relatedness between sentences. In the end, the system performed well for English and obtained a mean 0.7216 Pearson correlation, which ranked 33^{th} among 74. However, it performed less adequately for Spanish, obtaining only a mean 0.5158, which ranked 9^{th} out of 17. Witch regards the system submitted to the DSL shared task 2015, our system ranked 2^{nd} (out of 9 teams) on one of the test sets and 4^{th} (out of 7 teams) on the other. In detail, it achieved an overall accuracy of 95.24% and 92.77%, respectively. Automatic language identification is often the first processing stage

of many NLP applications and pipelines. Accordingly, this research and resulted code are of extreme importance for building better compilation tools capable of automatically discriminate between similar languages and language varieties.

Chapter 6

Conclusion

"Many a man has finally succeeded
only because he has failed after repeated efforts.
If he had never met defeat
he would never have known any great victory."

—Orison Swett Marden

The research described in this thesis is an answer towards our initial goals, which aimed at exploiting and developing new technologies and methods to better ascertain professionals and laypersons on compiling and managing multilingual corpora and terminology. In detail, this work allowed us to test and generate theories about the needs of professionals translators and interpreters, as well as ordinary people regarding the technologies they use, understand whether they are satisfied with existing technologies, suggest evaluation methods for those, identify problems that require more thorough research, as well as possible ways to improve existing tools and methodologies. The main findings of this work are summarised and discussed below, followed by suggestions for future research.

The work reported in this thesis can be clustered into three main Research Questions (RQ), which were carefully formulated in the introductory section. The way we started approaching the first RQ was by doing an extensive analysis on the existing comparable compilation tools on the market. Our findings showed that non of the analysed tools have a native bi- or multilingual comparable corpora compilation option or allow to use more than one Boolean operator when creating search query strings. After a careful analysis of their limitations and strengths, we decided to built a new open-source web-based multilingual comparable corpora prototype, named iCompileCorpora. iCompileCorpora[132] not only overcomes various spotted usability problems, limitations and performance issues, but also improves the current compilation process in terms of its flexibility and robustness. By a way of example it offers the possibility to compile mono-, bi- or even multilingual comparable corpora from the Web. Although there is always room for improvement and there are a lot of aspects we would like to either add or improve, we believe that solid steps have been taken in the right direction by showing that is possible to take advantage of the current technologies and build a simple, yet robust multilingual comparable compilation tool that is intuitive and easy-to-use by both professionals and laypersons. Another web tool build in this scope was the SCleaner[133]. When copying and pasting from a PDF file, users can find various formatting problems regarding extra white spaces, tabulations, sentence boundaries delimitations, amongst other issues. SCleaner automatic removes extra tabs and white spaces, and splits sentences in the right place automatically. Despite its simplicity, SCleaner is a very handy tool when dealing with unformatted text as it can speed-up this tedious and time-consuming task of formatting documents.

The second goal of this research focused on identifying the right variables that could be used to assess Terminology Management Tools (TMT) and Terminology Extraction Tools (TET). As a result we suggested standardised scoring systems that can be easily customised and, thus used by the users while comparing or simple determining the most adequate tool for a specific task in hand. An interesting finding from this exercise was the realisation of the scarcity of interpreting tools available on the market. Unlike translators, for whom a myriad of computer-assisted tools are available, interpreters have not benefited from the same level of automation or innovation, ending up using non or tools primarily designed for translation purposes. Moreover, various surveys reported that the existing type of tools both TMS and TET do not fulfil all interpreters' and translators' needs. Accordingly, the next step in the right direction could be to gather detailed information to better ascertain

[132] `https://icompilecorpora.herokuapp.com/home`
[133] `http://www.lexytrad.es/scleaner/index.php`

translators' and interpreters' technology awareness and real needs in order to design new tools or improve existing ones. Above all, the biggest surprise was to realise that there is nearly no advances in the TMT and TET stack of technology in the recent years. The current tools either have not released a new version in years or those that have, reported small improvements.

The third and last goal of this research mainly focused on exploring various methods capable of helping users accessing comparable corpora. As a first step, we focused on building PreProcessor[134], a program that offers a variety of morphosyntatic options to process and annotate raw textual data by taking advantage of the best known open-source libraries on the market. Then, we combined various Natural Language Processing (NLP) methods and Distributional Similarity Measures (DSM) into a program named STSModule[135]. This program allows us to assess semantic similarity between both sentences and documents in English. Finally, we proposed a simple, yet efficient methodology capable of assessing and ranking comparable documents according to their internal degree of similarity, which resulted in a third program called DSMModule[136]. This program not only can help the user to have a better idea about the quality of the documents in the corpus but also can help deciding which documents should belong or be removed from it. The next step would be the integration of these three programs in a compilation tool or even build an interface for it so non technical people could use it.

Although various ideas for future improvements have been already discussed through this thesis, we would like to give the reader a broader idea how strongly connected they are and how could they be tackled in the future. A consequent issue regarding the lake of good software for various computational linguistic tasks, like corpora compilation or terminology management and extraction is that, we spend too much time looking for unexciting or not reliably and complicated tools. Moreover, their creators are either companies interested in profit or research groups without conditions for long-term support. Although we can see some effort from both parties, there still is a long role to fulfil translators', interpreters' and laypersons' demands. Having this in mind and considering all the contributions reported in this thesis, we believe we did the first steps in the direction of building better software and methods capable of helping users assessing linguistic resources. If this project continues in the future, we would firstly focus on further improve iCompileCorpora by adding the option to compile parallel corpora. Then, we would direct our efforts on the second part of the compilation process, i.e. the managing part by building a second web-based application capable of managing sentences, documents and corpora (i.e. make possible to edit, copy and paste sentences and documents from and to documents and corpora respectively, as well as to manage corpora into domains and sub-domains); measuring the similarity between sentences, documents and corpora; and exploring the representativeness of the corpora. Finally, a third application would be responsible for exploring corpora, which would offer a set of concordance features, such as the ability to search for words in context, automatically extract the most frequent words and multiword units, amongst other features. All these three web applications could then be incorporated into a web

[134] https://github.com/hpcosta/PreProcessor
[135] https://github.com/hpcosta/STSModule
[136] https://github.com/hpcosta/DSMModule

platform. As a result, this platform would be the first one to offer the three compilation pillars, compilation, management and exploration. We are aware that there is plenty of work to do, but we sincerely hope that this project has the deserved continuity. However, there are unfortunately aspects that do not depend solely on us.

6.1 Conclusions in Spanish

La investigación que se aborda en la presente tesis doctoral surge para dar respuesta a los objetivos inicialmente establecidos, cuyo denominador común gira en torno a la idea de explotar y desarrollar nuevos métodos y tecnologías que asistan tanto a los profesionales como a los usuarios no especializados en la materia de compilación y gestión de corpus y terminología multilingüe. De esta forma, el presente estudio ha permitido el análisis e implementación de teorías sobre las necesidades de los traductores e intérpretes profesionales y de personas no especializas en la materia con respecto a las tecnologías utilizadas. Concretamente, se ha evaluado el grado de satisfacción con las tecnologías existentes, lo que ha permitido la propuesta de nuevos métodos de evaluación e identificación de problemas a fin de mejorar las herramientas y metodologías actuales. A continuación, se exponen sucintamente y se discuten las principales conclusiones de este trabajo, seguidas de posibles líneas de investigación futuras:

El trabajo presentado en la tesis se estructura en torno a las tres principales preguntas de investigación detalladamente formuladas en el apartado introductorio (cfr. Introducción). Así, para abordar la primera de ellas, se han analizado exhaustivamente las herramientas de compilación de corpus comparables existentes en el mercado. Nuestros hallazgos mostraron que, de las herramientas analizadas, ninguna de ellas cuenta con la funcionalidad de compilación de más de un corpus comparable a la vez, ni permite la utilización de más de un operador booleano en la creación de cadenas de consulta de búsqueda. Seguidamente, y una vez analizadas sus limitaciones y fortalezas, nos planteamos la implementación de un prototipo de gestión de corpus comparable multilingüe basado en web de código abierto, llamado *iCompileCorpora*[137]. De esta forma, *iCompileCorpora* pretende dar solución, tanto a los problemas detectados referentes a la usabilidad, como a las limitaciones y problemas de rendimiento, a la par que optimiza el proceso de compilación de corpus en términos de flexibilidad y confiabilidad. A modo de ejemplo, *iCompileCorpora* ofrece la opción de compilar corpus comparables mono-, bi- e incluso multilingües virtuales a través de textos descargados de la red Internet. Aunque aún existen aspectos que se podrían incluir o mejorar, consideramos que se han dado los primeros pasos en la dirección adecuada en tanto que se ha puesto de manifiesto la posibilidad de aplicar las virtudes de las tecnologías actuales para la implementación de una herramienta de compilación comparable, multilingüe, simple, fiable, intuitiva y de fácil manejo, tanto para profesionales como para usuarios no especialistas. En la misma línea, otra de las herramientas informáticas implementadas ha sido *SCleaner*[138] que, a pesar de su simplicidad, tiene como objetivo dar solución a los pequeños problemas derivados del proceso de dar formato a un texto copiado de un archivo en formato .pdf. Concretamente, *SCleaner* elimina, de manera automática, pequeñas erratas como tabulaciones y espacios adicionales e, incluso, es capaz de dividir las oraciones correctamente.

El segundo objetivo de esta investigación se centró en identificar los parámetros que podrían utilizarse para analizar las características tanto de las herramientas de gestión como de extracción terminológica. Seguidamente, se ha propuesto un

[137] `https://icompilecorpora.herokuapp.com/home`
[138] `http://www.lexytrad.es/scleaner/index.php`

sistema de evaluación estandarizado, fácilmente personalizable, que permitirá a los usuarios a comparar o determinar la herramienta más adecuada a la hora de abordar una determinada tarea. Por su parte, la carencia de herramientas de interpretación disponibles en el mercado ha sido uno de los resultados más interesantes dimanantes del mencionado análisis. De este modo, a diferencia de los traductores, para quienes sí existe una gran cantidad de herramientas informáticas asistidas, los intérpretes no se han beneficiado al mismo nivel de la automatización y la innovación, por lo que han terminado empleando herramientas diseñadas, principalmente, para fines traductológicos. A esto se unen los resultados de varias encuestas que han revelado los sistemas de gestión de traducciones y las herramientas de gestión de terminología, de manera general, no satisfacen las necesidades profesionales de los traductores e intérpretes. En consecuencia, el siguiente paso consiste en determinar exhaustivamente el conocimiento tecnológico y los requerimientos de los profesionales en el ámbito de la traducción y de la interpretación para el posterior diseño de nuevas herramientas o la mejora de las ya existentes. Uno de los hallazgos más sorprendente consistió en descubrir que apenas se habían producido avances científicos en las herramientas, tanto de gestión como de extracción terminológica o, si se había producido dicha implementación, ha supuesto la introducción de pequeñas mejoras con escasa relevancia.

El tercer y último objetivo de esta investigación está dedicado, principalmente, a explorar las diferentes metodologías que asisten a los usuarios en la compilación de corpus comparables. Para ello, en primer lugar, nos hemos enfocado en la creación de *PreProcessor*[139], un programa que ofrece variedad de funciones morfosintácticas para anotar datos de textos sin procesar, aprovechando las bibliotecas de código abierto más conocidas en el mercado. A continuación, se combinaron diversos métodos relacionados con el PLN y con las medidas de similitud distributiva y se obtuvo como resultado el programa *STSModule*[140], que permite evaluar la similitud semántica entre oraciones y los documentos en inglés. Finalmente, se ha propuesto una metodología sencilla, pero eficiente, capaz de evaluar y clasificar documentos comparables de acuerdo a su grado interno de similitud, lo que ha resultado en un tercer programa llamado *DSMModule*[141]. El programa *DSMModule* tiene como finalidad principal ayudar a determinar adecuadamente la representatividad cualitativa de los documentos que integran el corpus y, por lo tanto, permite discernir entre aquellas muestras que debe formar parte del mismo y las que deben eliminarse. El siguiente paso consistiría en la integración de estos tres programas en una herramienta de compilación, e incluso, en la construcción de una interfaz para personas no especializadas en la materia.

Con la realización de la presente tesis doctoral se han llevado a debate varias ideas para futuras mejoras, lo que pone en relieve cuan intrínsecamente estas están conectadas y cómo podrían abordarse en el futuro. El principal problema que puede inferirse con respecto a la carencia de un software de calidad para abordar las diversas tareas lingüísticas computacionales, como la compilación de un corpus o la gestión y extracción de terminología, es la excesiva demora que provoca pasar demasiado tiempo buscando herramientas poco dinámicas, no fiables y complejas. Además, sus creadores suelen ser empresas interesadas fundamentalmente en el

[139] `https://github.com/hpcosta/PreProcessor`
[140] `https://github.com/hpcosta/STSModule`
[141] `https://github.com/hpcosta/DSMModule`

beneficio económico o grupos de investigación que, aunque muestran un gran interés por satisfacer las demandas de los traductores o intérpretes, no cuentan con apoyo económico a largo plazo. Así, teniendo en cuenta estas cuestiones, y considerando todas las contribuciones introducidas en esta tesis doctoral, consideramos que hemos dado los primeros pasos hacia la construcción de mejores herramientas y métodos capaces de ayudar a los usuarios que emplean recursos lingüísticos, con especial referencia a los traductores e intérpretes. Una de las posibles líneas de investigación futuras que nos proponemos es la mejora de la herramienta informática *iCompileCorpora*, agregando la opción de gestión corpus paralelos. A continuación, centraríamos nuestros esfuerzos en la segunda parte del proceso de compilación, a saber, la parte de gestión mediante la creación de una segunda aplicación basada en una web que permita gestionar oraciones, documentos y corpus (es decir, hacer posible editar, copiar y pegar oraciones y documentos y corpus; y explorando la representatividad de los corpus). Finalmente, una tercera aplicación se encargaría de explorar los corpus, que brindaría un conjunto de características de concordancia, como la posibilidad de buscar palabras en contexto, extraer automáticamente las palabras más frecuentes y unidades terminológicas, entre otras funciones. Estas tres aplicaciones web, a su vez, podrían volcarse y aunarse a través de una plataforma en la red Internet. Como resultado, sería la primera plataforma en ofrecer los tres pilares principales en los que se apoya el proceso de compilación, a saber, compilación, gestión y explotación.

References

Abaitua, J. (2002). Tratamiento de corpora bilingües. *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita. Barcelona: Fundación Duques de Soria-Edicions Universitat de Barcelona (Manuals UB, 53)*, pages 61–90.

Agirre, E., Banea, C., Cardic, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. (2014). SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *8th Int. Workshop on Semantic Evaluation*, SemEval'14, pages 81–91, Dublin, Ireland. ACL and Dublin City University.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., and Wiebe, J. (2015). SEM 2013 shared task: Semantic Textual Similarity. In *9th Int. Workshop on Semantic Evaluation*, SemEval'15, pages 252–263, Denver, Colorado, USA. ACL.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *10th Int. Workshop on Semantic Evaluation*, SemEval'16, pages 497–511, San Diego, CA, USA. ACL.

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. (2012). SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *1st Joint Conf. on Lexical and Computational Semantics (*SEM): Proc. of the Main Conf. and the Shared Task*, pages 385–393, Montréal, Canada.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). SEM 2013 shared task: Semantic Textual Similarity. In *2nd Joint Conf. on Lexical and Computational Semantics (*SEM), Volume 1: Proc. of the Main Conf. and the Shared Task*, pages 32–43, Atlanta, Georgia, USA.

Anthony, L. (2014). AntConc (Version 3.4.3) Machintosh OS X. Waseda University. Tokyo, Japan. Available from `http://www.laurenceanthony.net`.

Arce Romeral, L. and Seghiri, M. (2018a). Compilation of an ad hoc corpus for extracting glossaries in the interpretation classroom. *Current Trends in Translation Teaching and Learning E*, pages 1–46.

Arce Romeral, L. and Seghiri, M. (2018b). Determinación de la representatividad cualitativa y cuantitativa de un corpus virtual de contratos de compraventa de viviendas (inglés-español). In *Traducción literaria y discursos traductológicos especializados*, pages 309–330. Peter Lang, Amsterdam, Netherlands.

Aston, G. (2016). (*In Press*) How corpora can help the interpreter walk the tightrope. In *Corpus-based Approaches to Translation and Interpreting: from theory to applications*. Peter Lang.

Atkins, S., Clear, J., and Ostler, N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing*, 7(1):1–16.

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In M. Baker, M. F. and Tognini-Bonelli, E., editors, *Text and Technology. In Honour of John Sinclair*, pages 233–250, Philadelphia/Amsterdam. John Benjamins Publishing Company.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. Bloomsbury Discourse. Bloomsbury Academic.

Banchs, R. E., D'Haro, L. F., and Li, H. (2015). Adequacy-fluency Metrics: Evaluating MT in the Continuous Space Model Framework. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(3):472–482.

Baroni, M. (2013). Composition in Distributional Semantics. *Language and Linguistics Compass*, 7(10):511–522.

Baroni, M., Bernardi, R., and Zamparelli, R. (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9(1).

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *4th Int. Conf. on Language Resources and Evaluation*, LREC'04, pages 1313–1316.

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *11th Annual Conf. of the European Association for Machine Translation*, EAMT'06, pages 247–252, Oslo, Norway. The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway).

Bekavac, B., Osenova, P., Simov, K., and Tadić, M. (2004). Making Monolingual Corpora Comparable: a Case Study of Bulgarian and Croatian. In *4th Language Resources and Evaluation Conf.*, LREC'04, pages 1187–1190, Lisbon, Portugal.

Bendazzoli, C. and Sandrelli, A. (2009). Corpus-based Interpreting Studies: Early Work and Future Prospects. *Tradumàtica*, 7.

Bergenholtz, H. and Tarp, S. (2003). Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. *Hermes, Journal of Linguistics*, 31:171–196.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge, UK.

Biber, D. (1990). Methodological Issues Regarding Corpus-based Analyses of Linguistic Variations. *Literary and Linguistic Computing*, 5(4):257–269.

Biber, D. (1993). Representativiness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press, Cambridge, UK.

Bilgen, B. (2009). *Investigating Terminology Management for Conference Interpreters*. MA dissertation, University of Otawa, Otawa, Canada.

Bilgen, B. (2011). *Investigating Terminology Management for Conference Interpreters: A User-oriented Study*. LAP Lambert Academic Publishing.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2013). Using WordNet and Semantic Similarity for Bilingual Terminology Mining from Comparable Corpora. In *6th Workshop on Building and Using Comparable Corpora (BUCC'13)*, pages 16–23, Sofia, Bulgaria.

Bowker, L. (2002). *Computer-aided Translation Technology: A Practical Introduction*. Didactics of translation series. University of Ottawa Press.

Bowker, L. and Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Braschler, M. and Scäuble, P. (1998). Multilingual Information Retrieval Based on Document Alignment Techniques. In *2nd European Conf. on Research and Advanced Technology for Digital Libraries*, pages 183–197. Springer.

Cencini, M. (2002). On the Importance of an Encoding Standard for Corpus-Based Interpreting Studies. Extending the TEI Scheme. *inTRAlinea, Special Issue CULT2K*.

Chen, J. and Nie, J.-Y. (2000). Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. In *6th Conf. on Applied Natural Language Processing*, pages 21–28.

Corley, C. and Mihalcea, R. (2005). Measuring the Semantic Similarity of Texts. In *ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE'05, pages 13–18, Stroudsburg, PA, USA.

Corpas Pastor, G. (2001). Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, 5(1):155–184.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Studien Zur Romanischen Sprachwissenschaft Und Interkulturellen Kommunikation. Peter Lang Pub Incorporated.

Corpas Pastor, G. (2018). Tools for Interpreters: the Challenges that Lie Ahead. *Current Trends in Translation Teaching and Learning E*, pages 138–184.

Corpas Pastor, G. and Durán Muñoz, I., editors (2017). *Trends in E-Tools and Resources for Translators and Interpreters*, volume 45 of *Approaches to Translation Studies*. Brill, Leiden, Netherlands.

Corpas Pastor, G. and Seghiri, M. (2007a). Determinación del Umbral de Representatividad de un Corpus mediante el Algoritmo N-Cor. *SEPLN: Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 39:165–172.

Corpas Pastor, G. and Seghiri, M. (2007b). Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness. *Translation Journal*, 11(3):19.

Corpas Pastor, G. and Seghiri, M. (2009). Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). In Beeby, A., Inés, P., and Sánchez-Gijón, P., editors, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

Corpas Pastor, G. and Seghiri, M., editors (2016). *Corpus-based Approaches to Translation and Interpreting: From Theory to Applications*. Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation, 106. Peter Lang, Frankfurt, Germany.

Costa, H. (2010). Automatic Extraction and Validation of Lexical Ontologies from text. Master's thesis, University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering, Coimbra, Portugal.

Costa, H. (2015). Assessing Comparable Corpora through Distributional Similarity Measures. In *EXPERT Scientific and Technological Workshop*, pages 23–32, Malaga, Spain. Tradulex.

Costa, H., Béchara, H., Taslimipoor, S., Gupta, R., Orăsan, C., Corpas Pastor, G., and Mitkov, R. (2015a). MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. In *9th Int. Workshop on Semantic Evaluation*, SemEval'15, pages 96–101, Denver, Colorado. ACL.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014a). A comparative User Evaluation of Terminology Management Tools for Interpreters. In *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, pages 68–76, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014b). Technology-assisted Interpreting. *MultiLingual #143*, 25(3):27–32.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2015b). An Interpreters' Guide to Selecting Terminology Management Tools. In *NATO Conf. on Terminology Management*, Brussels, Belgium.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2017). Assessing Terminology Management Systems for Interpreters. In Corpas Pastor, G. and Durán Muñoz, I., editors, *Trends in E-Tools and Resources for Translators and Interpreters*, volume 45, pages 57–84. Brill, Leiden, Netherlands.

Costa, H., Corpas Pastor, G., and Mitkov, R. (2015c). Measuring the Relatedness between Documents in Comparable Corpora. In *11th Int. Conf. on Terminology and Artificial Intelligence*, TIA'15, pages 29–37, Granada, Spain.

Costa, H., Corpas Pastor, G., and Seghiri, M. (2014c). iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, London, UK.

Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015d). iCorpora: Compiling, Managing and Exploring Multilingual Data. In *7th Int. Conf. of the Iberian Association of Translation and Interpreting Studies*, AIETI, pages 74–76, Malaga, Spain.

Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015e). Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora. In *New Horizons in Translation and Interpreting Studies (Full papers)*, pages 133–141, Genebra, Switzerland. Tradulex.

Costa, H., Durán Muñoz, I., Corpas Pastor, G., and Mitkov, R. (2016a). Compilação de Corpos Comparáveis Especializados: Devemos sempre confiar nas Ferramentas de Compilação Semi-automáticas? *Linguamática*, 8(1):3–18.

Costa, H., Gonçalo Oliveira, H., and Gomes, P. (2010). The Impact of Distributional Metrics in the Quality of Relational Triples. In *19th European Conf. on Artificial Intelligence, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, ECAI'10, pages 23–29, Lisbon, Portugal.

Costa, H., Gonçalo Oliveira, H., and Gomes, P. (2011). Using the Web to Validate Lexico-Semantic Relations. In *15th Portuguese Conf. on Artificial Intelligence*, volume 7026 of *EPIA'11*, pages 597–609, Lisbon, Portugal. Springer.

Costa, H., Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2016b). Nine terminology extraction Tools: Are they useful for translators? *MultiLingual #159*, 27(3).

de Groc, C. (2011). Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. In *IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology - Volume 1*, WI-IAT'11, pages 497–498, Lyon, France. IEEE Computer Society.

Defrancq, B. (2016). (*In Press*) Well, interpreters... a corpus-based study of a pragmatic particle used by simultaneous interpreters. In *Corpus-based Approaches to Translation and Interpreting: from theory to applications*. Peter Lang.

Durán Muñoz, I. (2012). Meeting Translators' Needs: Translation-oriented Terminological Management and Applications. *The Journal of Specialised Translation*, 18:77–92.

EAGLES (1994). Corpus Typology: A framework for classification. Tech Report N.2.1 written by John M. Sinclair, EAGLES Document 080294, Corpus Linguistics Group, Universidad de Birmingham, UK.

EAGLES (1996a). Evaluation of Natural Language Processing Systems. Technical report, EAGLES Document EAG-EWG-PR.2. `http://www.issco.unige.ch/en/research/projects/ewg96/ewg96.html`.

EAGLES (1996b). Preliminary Recommendations on Corpus Typology. Technical report, EAGLES Document EAG-TCWG-CTYP/P. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

EAGLES (1996c). Text Corpora Working Group Reading Guide. Technical report, EAGLES Document EAG-TCWG-FR-2. `http://www.ilc.cnr.it/EAGLES/corpintr/corpintr.html`.

Eisele, A. and Xu, J. (2010). Improving Machine Translation Performance Using Comparable Corpora. In *3rd Workshop on Building and Using Comparable Corpora (BUCC'10)*, pages 35–41, La Valletta, Malta.

Faber, P. (2015). Frames as a framework for terminology. In *Handbook of Terminology*, pages 14–33. John Benjamins Publishing Company.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Fantinuoli, C. and Zanettin, F., editors (2015). *New directions in corpus-based translation studies*. Translation and Multilingual Natural Language Processing 1. Language Science Press, Berlin, Germany.

Firth, J. R. (1935). The Technique of Semantics. In *Transactions of the Philological Society*, pages 36–72.

Flowerdale, L. (2004). The argument for using English specialised corpora to un academic and professional language. In Connor, U. and Upton, T., editors, *Discourse In The Professions: Perspectives From Corpus Linguistics*, pages 11–33, Amsterdam/Philadelphia. John Benjamins.

Fung, P. and Cheung, P. (2004). Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *Conf. on Empirical Methods in Natural Language Processing*, EMNLP'04, pages 57–63, Barcelona, Spain.

Gamallo, P. and González López, I. (2010). Wikipedia as a multilingual source of comparable corpora. In *3rd Workshop on Building and Using Comparable Corpora (BUCC'10)*, pages 21–25, Valletta, Malta.

Gebre, B. G., Zampieri, M., Wittenburg, P., and Heskens, T. (2013). Improving Native Language Identification with TF-IDF Weighting. In *8th Workshop on Innovative Use of NLP for Building Educational Applications*.

Gil-Berrozpe, J. C. and Faber, P. (2016). Refining Hyponymy in a Terminological Knowledge Base. In *2nd Workshop on Language and Ontology (LangOnto2) & Terminology and Knowledge Structures (TermiKS) at the 10th Language Resources and Evaluation Conference (LREC'16)*, pages 8–15, Portorož, Slovenia.

Goeuriot, L., Morin, E., and Daille, B. (2009). Compilation of Specialized Comparable Corpora in French and Japanese. In *2nd Workshop on Building and Using Comparable Corpora (BUCC'09)*, pages 55–63, Singapore. ACL.

Gonçalo Oliveira, H., Costa, H., and Gomes, P. (2010). Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. In *INForum 2010 - II Simpósio de Informática, Track on Gestão e Tratamento de Informação*, INForum'10, pages 537–548, Braga, Portugal.

Goutte, C., Léger, S., and Carpuat, M. (2014). The NRC System for Discriminating Similar Languages. In *VarDial Workshop*, Dublin, Ireland.

Grishman, R. (1997). Information Extraction: Techniques and Challenges. In *Int. Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, SCIE'97, pages 10–27, London, UK. Springer.

Gupta, R., Bechara, H., El Maarouf, I., and Orăsan, C. (2014). UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment. In *8th Int. Workshop on Semantic Evaluation (SemEval'14)*, pages 785–789, Dublin, Ireland. ACL and Dublin City University.

Gutiérrez Florido, R., Corpas Pastor, G., and Seghiri, M. (2013). Using semi-automatic compiled corpora for medical terminology and vocabulary building in the healthcare domain. In *10th Int. Conf. on Terminology and Artificial Intelligence (TIA'13), Workshop on Optimizing Understanding in Multilingual Hospital Encounters*, Paris, France.

Harris, Z. (1970). Distributional Structure. In *Papers in Structural and Transformational Linguistics*, pages 775–794. D. Reidel Publishing Company, Dordrecht, Holland.

Hashemi, H. B., Shakery, A., and Faili, H. (2010). Creating a Persian-English Comparable Corpus. In *2010 Int. Conf. on Multilingual and Multimodal Information Access Evaluation: Cross-language Evaluation Forum*, CLEF'10, pages 27–39. Springer.

Hazem, A. and Morin, E. (2013). A Comparison of Smoothing Techniques for Bilingual Lexicon Extraction from Comparable Corpora. In *6th Workshop on Building and Using Comparable Corpora (BUCC'13)*, pages 24–33, Sofia, Bulgaria.

Ibrahimov, O., Sethi, I., and Dimitrova, N. (2002). The Performance Analysis of a Chi-square Similarity Measure for Topic Related Clustering of Noisy Transcripts. In *16th Int. Conf. on Pattern Recognition*, volume 4, pages 285–288. IEEE Computer Society.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European Conf. on Machine Learning*, ECML'98, pages 137–142. Springer.

Johansson, S. and Oksefjell, S. (1998). *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Rodopi.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., NJ, USA.

Kilgarriff, A. (2001). Comparing Corpora. *Int. Journal of Corpus Linguistics*, 6(1):1–37.

Kilgarriff, A. (2010). Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project. In *3rd Workshop on Building and Using Comparable Corpora (BUCC'10)*, pages 1–5, La Valletta, Malta.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine. In *11th EURALEX Int. Congress*, pages 105–116, Lorient, France.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit*.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Koehn, P., Och, F., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *NAACL-HLT - Volume 1*, pages 48–54. ACL.

Köhler, R. (2013). Statistical Comparability: Methodological Caveats. In Sharoff, S., Rapp, R., Zweigenbaum, P., and Fung, P., editors, *Building and Using Comparable Corpora*, pages 77–91. Springer.

Kotani, K. and Yoshimi, T. (2015). Application of a Corpus to Identify Gaps between English Learners and Native Speakers. In *8th Workshop on Building and Using Comparable Corpora (BUCC'15)*, pages 38–42, Beijing, China. ACL.

Lavid López, J. (2005). *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*. Catédra, Madrid, Spain.

Laviosa, S. (2016). (*In Press*) Corpus-based translanguaging for translation education. In *Corpus-based Approaches to Translation and Interpreting: from theory to applications*. Peter Lang.

Lee, L. (1999). Measures of Distributional Similarity. In *37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL'99, pages 25–32. ACL.

Leturia, I., Vicente, I. S., and Saralegi, X. (2009). Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet. In *5th Int. Web as Corpus Workshop*, WAC5, pages 53–61, Donostia/San Sebastian, Spain.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL-HLT - Volume 1*, pages 71–78.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *15th Int. Conf. on Machine Learning*, ICML'98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann.

Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. In *ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP'02, pages 62–69. ACL.

Lüdeling, A. and Kytö, M. (2008). *Corpus linguistics: an international handbook*. Number v. 1 in Handbücher zur Sprach- und Kommunikationswissenschaft. W. de Gruyter.

Lui, M. and Baldwin, T. (2011). Cross-domain Feature Selection for Language Identification. In *5th Int. Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Ma, X. and Liberman, M. (1999). BITS: A Method for Bilingual Text Search over the Web. In *Machine Translation Summit VII*.

Ma, Y., He, Y., Way, A., and van Genabith, J. (2011). Consistent translation using discriminative learning: a translation memory-inspired approach. In *49th Annual Meeting of the ACL: Human Language Technologies - Volume 1*, pages 1239–1248, Portland, Oregon, USA.

Maia, B. (2003). What are comparable corpora? In Hansen-Schirra, S. and Neumann, S., editors, *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*, pages 27–34, Lancaster, UK.

McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. Routledge.

Meunier, J.-L. and Dymetman, M. (2014). Extended Translation Memories for Multilingual Document Authoring. In *7th Workshop on Building and Using Comparable Corpora (BUCC'14)*, pages 28–37, Reykjavik, Iceland. ACL.

Mitkov, R. (2016). The Name of the Game is Comparable Corpora. In *9th Workshop on Building and Using Comparable Corpora (BUCC'16)*, pages 1–2, Portorož, Slovenia. ELDA.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. In *45th Annual Meeting of the ACL*, pages 664–671, Prague, Czech Republic. ACL.

Moser-Mercer, B. (1992). Banking on Terminology: Conference Interpreters in the Electronic Age. *Meta: Translators' Journal*, 37(3):507–522.

Munteanu, D. S. and Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

Navigli, R. and Paolo Ponzetto, S. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Niculae, V., Zampieri, M., Dinu, L. P., and Ciobanu, A. M. (2014). Temporal Text Ranking and Automatic Dating of Texts. In *14th Conf. of the European Chapter of the Association for Computational Linguistics*, EACL'14, pages 17–21, Gothenburg, Sweden. ACL.

Och, F. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *41st Annual Meeting on ACL - Volume 1*, ACL'03, pages 160–167. ACL.

Och, F. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Orăsan, C., Cattelan, A., Corpas Pastor, G., van Genabith, J., Herranz, M., Arevalillo, J. J., Liu, Q., Sima'an, K., and Specia, L. (2015). The EXPERT Project: Advancing the State of the Art in Hybrid Translation Technologies. In *Translating and the Computer 37 - AsLing*, London, UK.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

Pérez-Pérez, P. (2018). The Use of a Corpus Management Tool for the Preparation of Interpreting Assignments: A Case Study. *The Int. Journal for Translation and Interpreting Research*, 10(1):137–151.

Pilehvar, M. T., Jurgens, D., and Navigli, R. (2013). Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In *51st Annual Meeting of the ACL - Volume 1*, pages 1341–1351, Sofia, Bulgaria. ACL.

Poibeau, T. (2017). *Machine Translation*. MIT Press.

Quirk, R. (1992). On Corpus Principles and Design. In *Directions in Corpus Linguistics*, Nobel Symposium 82, pages 457–469, Stockholm, Sweden. Walter de Gruyter.

Rafajlovska, A. (2013). *Natural Language Processing Approach for Macedonian-French and Macedonian-English Interpreting based on Oral Sociopolitical Corpora*. Master Thesis, Université de Franche-Comté, France and Universidade do Algarve, Portugal.

Rapp, R., Sharoff, S., and Zeigenbaum, P. (2016). Special Issue on using comparable corpora for Machine Translation. *Journal of Natural Language Engineering*, 22(4).

Rayson, P., Leech, G., and Hodges, M. (1997). Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. *Int. Journal of Corpus Linguistics*, 2(1):133–152.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal Of Artificial Intelligence Research (JAIR)*, 11:95–130.

Rodríguez, N. and Schnell, B. (2009). A Look at Terminology Adapted to the Requirements of Interpretation. *Language Update*, 6(1):21–27.

Salton, G. and Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Santos, D. and Rocha, P. (2001). Evaluating CETEMPúblico, a free resource for portuguese. In *39th Annual Meeting on Association for Computational Linguistics*, ACL'01, pages 450–457. ACL.

Saralegi, X., naki San Vicente, I., and Gurrutxaga, A. (2008). Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *1st Workshop on Building and Using Comparable Corpora (BUCC'08)*, pages 27–32, Marrakech, Morocco.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *ACL, SIGDAT Workshop*, pages 47–50, Dublin, Ireland.

Seghiri, M. (2006). *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad*. PhD Thesis, University of Malaga, Spain.

Seghiri, M. (2008). Creating virtual corpora step by step. In P. Sánchez Hernández, editor, *Researching and teaching specialized languages: new contexts, new challenges*. REBIUN.

Seghiri, M. (2011). Metodología protocolizada de compilación de un corpus de seguros de viajes: aspectos de diseño y representatividad. *Revista de Lingüística Teórica y Aplicada (RLA)*, 49(2):13–30.

Seghiri, M. (2015). Determinación de la representatividad cuantitativa de un corpus ad hoc bilingüe (inglés-español) de manuales de instrucciones generales de lectores electrónicos/ Establishing the quantitative representativiness of an E-Reader User's Guide ad hoc corpus (English-Spanish). In María Teresa Sánchez Nieto, editor, *Corpus-based Translation and Interpreting Studies: From description to application*, pages 125–146. Frank & Timme, Berlin, Germany.

Seghiri, M. (2016). Diseño de una plantilla electrónica de evaluación de sedes web científicas para la creación de recursos de enseñanza-aprendizaje (alemán-inglés-español)/Designing an evaluation template of scientific web sites for the implementation of teaching and learning materials (German-English-Spanish). *Educatio Siglo XXI*, 34(3):9–26.

Seghiri, M. (2017a). Corpus e interpretación biosanitaria: extracción terminológica basada en bitextos del campo de la Neurología para la fase documental del intérprete (Corpora and medical interpreting: terminology extraction based on bitexts for the interpreter's documentation process). *Volumen monográfico de interpretación en el ámbito biosanitario*, 46(18):123–132.

Seghiri, M. (2017b). Metodología de elaboración de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus para la traducción de manuales de instrucciones de televisores/ Creating a bilingual and bidirectional glossary (English-Spanish/Spanish-English) based on corpus for the translation of TV user manuals). *Babel*, 63(1):43–64.

Sharoff, S. (2007). Classifying Web corpora into domain and genre using automatic feature identification. In *Web as Corpus Workshop*, UCLouvain, Louvain-la-Neuve, Belgium.

Sharoff, S. (2010). Analysing similarities and differences between corpora. In *7th Conf. of Language Technologies (Jezikovne Tehnologije)*, pages 5–11, Ljubljana, Slovenia.

Sharoff, S. (2013). Measuring the Distance Between Comparable Corpora Between Languages. In Sharoff, S., Rapp, R., Zweigenbaum, P., and Fung, P., editors, *Building and Using Comparable Corpora*, pages 113–130. Springer.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42.

Skadiņa, I., Aker, A., Giouli, V., Tufis, D., Gaizauskas, R., Mieriņa, M., and Mastropavlos, N. (2010a). A Collection of Comparable Corpora for Under-resourced Languages. In *4th Int. Conf. Baltic HLT: The Baltic Perspective*, pages 161–168. IOS Press.

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiş, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M. L., and Pinnis, M. (2012). Collecting and Using Comparable Corpora for Statistical Machine Translation. In *8th Int. Conf. on Language Resources and Evaluation*, LREC'12, pages 438–445, Istanbul, Turkey.

Skadiņa, I., Vasiļjevs, A., Skadiņŝ, R., Gaizauskas, R., Tufiş, D., and Gornostay, T. (2010b). Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. In *3rd Workshop on Building and Using Comparable Corpora (BUCC'10)*, pages 6–14, Valletta, Malta.

Snover, M., Li, X., Lin, W.-P., Chen, Z., Tamang, S., Ge, M., Lee, A., Li, Q., Li, H., Anzaroot, S., and Ji, H. (2011). Cross-lingual Slot Filling from Comparable Corpora. In *4th Workshop on Building and Using Comparable Corpora (BUCC'11)*, pages 110–119, Portland, Oregon, USA.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Overview for the First Shared Task on Language Identification in Code-Switched Data. In *1st Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar.

Somers, H. (2003). Translation memory systems. In Somers, H., editor, *Computers and Translation: A translator's guide*, pages 31–49. John Benjamins.

Spohr, D. (2009). Towards a Multifunctional Electronic Dictionary Using a Metamodel of User Needs. In *eLexicography in the 21st century: New challenges, new applications*, Louvain-La-Neuve, Belgium. Presses Universitaires de Louvain.

Stolcke, A. (2002). SRILM - an Extensible Language Modeling Toolkit. In *7th Int. Conf. on Spoken Language Processing*, ICSLP'02, pages 901–904.

Straniero S., Falbo, C., editor (2012). *Breaking Ground in Corpus-based Interpreting Studies*. Peter Lang, Bern, Switzerland.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., and Keskustalo, H. (2007). Creating and Exploiting a Comparable Corpus in Cross-language Information Retrieval. *ACM Transactions on Information Systems (TOIS)*, 25(1).

Tan, L., Zampieri, M., Ljubešić, N., and Tiedemann, J. (2014). Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection. In *7th Workshop on Building and Using Comparable Corpora (BUCC'14)*, pages 6–10, Reykjavik, Iceland. ACL.

Tarp, S. (2008). *Lexicography in the Borderland Between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Lexicographica: Series maior. Walter de Gruyter, 1^{st} edition.

Taylor, C. (2008). What is *corpuslinguistics*? What the data says? *ICAME Journal*, 32(1):179–200.

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive-Approach.* John Benjamins Publishing Company.

Torruella, J. and Llisterri, J. (1999). Diseño de corpus textuales y orales. In *Filología e informática: Nuevas tecnologias en los estudios filológicos*, Seminario de Filología e Informática de la Universidad Autónoma de Barcelona y Ed. Milenio, pages 45–77, Barcelona, Spain. José Manuel Blecua, Gloria Clavería, Cárlos Sánchez y Joan Torruella.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *HLT-NAAC 2003*, pages 252–259, Edmonton, Canada. ACL.

Turney, P. D. (2001). Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. In *12^{th} European Conf. on Machine Learning*, EMCL'01, pages 491–502, London, UK. Springer.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Varantola, K. (2003). *Translators and Disposable Corpora*, pages 55–70. Saint Jerome Publishing.

Vela, M. and Tan, L. (2015). Predicting machine translation adequacy with document embeddings. In *10^{th} Workshop on Statistical Machine Translation*, pages 402–410, Lisbon, Portugal.

Zampieri, M. and Gebre, B. G. (2012). Automatic Identification of Language Varieties: The Case of Portuguese. In *KONVENS*, pages 233–237, Vienna, Austria.

Zampieri, M. and Gebre, B. G. (2014). VarClass: An Open Source Language Identification Tool for Language Varieties. In *9^{th} Int. Conf. on Language Resources and Evaluation*, LREC'14, Reykjavik, Iceland.

Zampieri, M., Gebre, B. G., Costa, H., and van Genabith, J. (2015). Comparing Approaches to the Identification of Similar Languages. In *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial'15), 2^{nd} Discriminating between Similar Languages Shared Task (DSL'15)*, page 7, Hissar, Bulgaria.

Zampieri, M., Gebre, B. G., and Diwersy, S. (2013). N-Gram Language Models and POS Distribution for the Identification of Spanish Varieties. In *Traitement Automatique des Langues Naturelles*, TALN'13, pages 580–587, Les Sables-d'Olonne, France.

Zampieri, M., Tan, L., Ljubešić, N., and Tiedemann, J. (2014). A Report on the DSL Shared Task 2014. In *VarDial Workshop*, pages 58–67.

Zanettin, F., Bernardini, S., and Stewart, D. (2003). *Corpora in Translator Education*. Manchester: St. Jerome Publishing.

Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2015). Translators' requirements for translation technologies: a user survey. In *New Horizons in Translation and Interpreting Studies (Full papers)*, pages 247–254, Geneva, Switzerland. Tradulex.

Zubiaga, A., San Vicente, I., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza, A., and Fresno, V. (2014). Overview of TweetLID: Tweet language identification at SEPLN 2014. In *Twitter Language Identification Workshop (tweetLID), XXX Conf. of the Spanish Society for Natural Language Processing (SEPLN'14)*, pages 1–11, Girona, Spain.

Appendix A

Publications

Costa et al. (2014b)

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014). **Technology-assisted Interpreting**. *MultiLingual* #143, 25(3):27-32

Technology-Assisted Interpreting

Hernani Costa*
University of Malaga
Malaga, Spain
hercos@uma.es

Gloria Corpas Pastor
University of Malaga
Malaga, Spain
gcorpas@uma.es

Isabel Durán Muñes
University of Malaga
Malaga, Spain
iduran@uma.es

Abstract

Unlike translators, for whom a myriad of computer-assisted tools are available, interpreters have not benefited from the same level of automation or innovation. Their work relies by and large on traditional or manual methods. The solutions tailored to the interpreters' needs are few and still far behind. Fortunately, there is a growing interest in developing tools addressed at interpreters as end users, although the number of these technology tools is still very low and they are not intended to cover all interpreters' needs.

1 Interpreting modes and opportunities for technology

The main categories of interpreting are simultaneous and consecutive interpreting, which refers to the mode of delivering the original message. In simultaneous interpreting, the target message is given at roughly the same time that the source message is produced, whereas in consecutive interpreting the interpreter waits until the speaker has finished before beginning the interpretation and takes notes in the meantime. Apart from these two main categories, we can also include a third one: liaison interpreting, which can be either simultaneous or consecutive. Liaison interpreters work in both directions for two parties, thus the languages being used become passive and active at the same time. Other common modes practiced are whispering interpreting, sight interpreting and sign language interpreting. Interpreting modes can be further classified according to the technical equipment used, the settings, the fields of expertise and topics. However, there is not yet a single, accepted classification. Relevant authors and reputable interpreting institutions such as ITI[1] or AIIC[2] have their own classifications. The list below comprises the most frequent interpreting modes encountered in industry literature and offered by company services. By no means is it intended to be exhaustive.

- *Whispered interpreting* (also *chuchotage*) is a subcategory of simultaneous interpreting whispered into the listener's ear for which no specialised equipment is required.
- *Conference interpreting* takes place in multilingual conferences and it can be either simultaneous or consecutive interpreting, depending on the capacity of the conference and on the technical equipment available.
- *Business interpreting* is a subcategory of liaison interpreting used for smaller groups or business meetings, visits to a foreign country, one-on-one interviews and so on.
- *Court interpreting* refers to interpreting services provided in a legal setting such as courts of law. It could be either simultaneous or consecutive, depending on the technical equipment and the audience.
- *Teleinterpreting* (also *remote interpreting*) is done through a remote or offsite interpreter via telephone (*over the phone interpreting, OPI*) or via video (*video remote interpreting,*

*Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement N° 317471.

[1] www.iti.org.uk
[2] www.aiic.net

VRI), especially in services related to community interpreting. It is mostly consecutive, but it can also be simultaneous.

- *Community interpreting* is another subcategory of liaison interpreting; its main aim is "*to enable people who are not fluent speakers of the official language(s) of the country to communicate with the providers of public services so as to facilitate full and equal access to legal, health, education, government, and social services*" (Roberts, 1994:127).

There is a manifold of possible interpreting and scenarios, and, therefore, any technology tools developed for interpreters should necessarily account for this fact. Most interpreting services (except for teleinterpreting) are on-site, meaning the clients are in the same place where the service takes place. This limits the possibilities to use a suite of tools to assist interpretation. To the best of our knowledge, such a system has not yet been developed. However, thanks to the development of smart phones, notebooks and tablets, interpreters have at their disposal some useful applications (see section 2).

The chances to develop tools for interpreters increase with regard to the preparation phase prior to any interpreting service, when interpreters need to acquire as much information and specialised knowledge as possible in order to get ready for their work. Once interpreters know the topic, the setting and all the features of the interpreting service, they can start compiling terminological resources such as glossaries, managing documents and so on. The correct management of these tools will usually mean better output. Another scenario prone to technology developments is training, where all kind of software and applications could be used to train interpreters at various stages and in different modes.

2 Technology tools for interpreters

Several tools and applications have been implemented to meet the needs in different interpreting contexts and modes. Even though some interpreters still store information and terminology on scraps of paper or excel spreadsheets, there are some specialised computer and mobile software that can be used to compile, store, manage and search within glossaries. They can typically be used to prepare an interpretation in consecutive interpreting or in a booth. Those applications are quite similar to the look-up terminology tools currently used by translators.In fact, some of them have been developed to cater to the needs of both translators and interpreters.

Intragloss[3] is a Mac OS X software created specifically to help interpreters when preparing for an event by allowing them to manage glossaries. This application can be simply defined as a glossary and document management tool created to help the interpreter prepare, use and merge different glossaries with preparation documents, in more than 150 different languages. It allows to import and export glossaries from and to Microsoft Word and Excel formats. Every glossary imported to or created in is assigned to a domain glossary, which contains all the glossaries from the sub-areas of knowledge, named 'assignments'. The creation of an assignment glossary can be done in two different ways: either by extracting (automatically or manually) all the terms from the domain glossary that appear in the documents, or by highlighting a term in the document, search for it on search sites (such as online glossaries, terminology databases, dictionaries and general Web pages) and adding the new translated term to the assignment glossary. The system allows for adding remarks, i.e. meta-information, to the glossary entries (see Fig. 1).

In short, Intragloss is an intuitive and easy-to-use tool that facilitates the interpreters' terminology management process by producing glossaries (imported or created *ad hoc*), by searching on several websites simultaneously and by highlighting all the terms in the documents that appear in the domain glossary. However, it is currently platform dependent and only works on Mac OS X platforms.

InterpretBank[4] is a simple terminology and knowledge management software tool designed both for interpreters and translators using Windows and Android. It helps to manage, learn and look up glossaries and term-related information. Due to its modular architecture, it can be used to guide the interpreter during the entire workflow process, starting from

[3] https://intragloss.com
[4] www.interpretbank.de

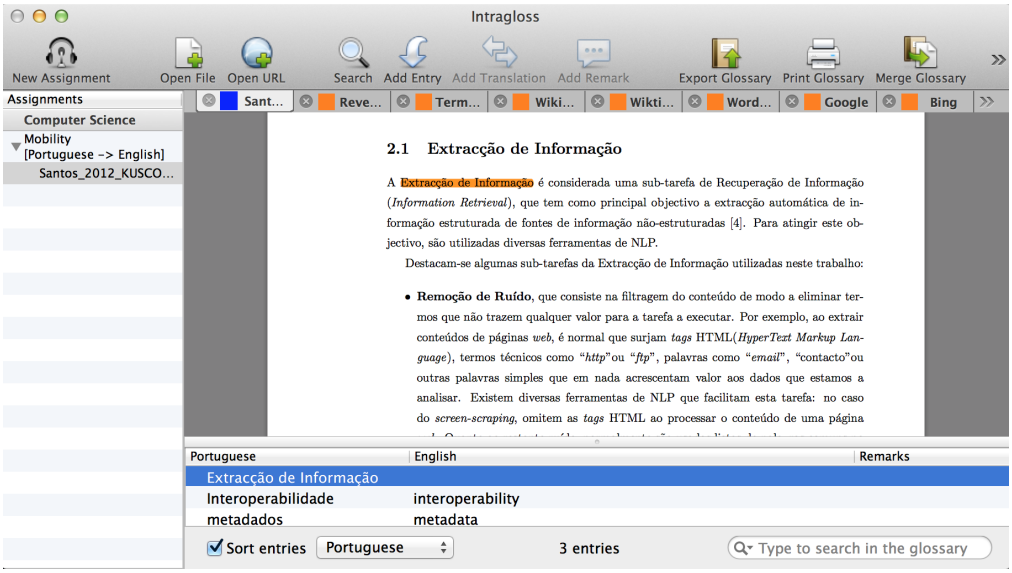


Figure 1: *Intragloss* screenshot.

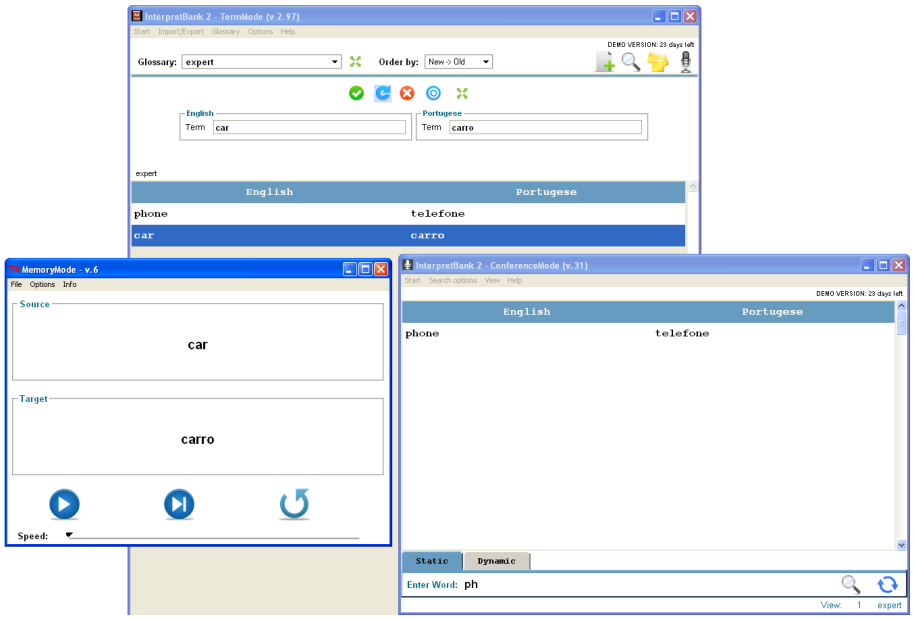


Figure 2: *InterpretBank* screenshot.

the creation and management of multilingual glossaries (TermMode), passing through the study of these glossaries (MemoryMode), and finally allowing the interpreter to look up terms while in a booth (ConferenceMode). See Fig. 2.

InterpretBank has also an Android version called InterpretBank Lite. This application is a simplified version of InterpretBank, specifically designed to access bi- or trilingual glossaries previously created with the desktop version. It is useful when working as a consecutive, community or liaison interpreter, when a quick look up at the terminology list is necessary.

InterpretBank has a user-friendly, intuitive and easy-to-use interface. It allows us to import and export glossaries in different formats (Microsoft Word, Microsoft Excel, simple text files, Android and TMEX) and automatically proposes translations to terms by taking advantage of online translation portal services. However, it is platform dependent (only works on Windows), it does not handle documents, only glossaries, and

it requires a commercial license.

Another user-friendly multi-lingual glossary management programme that can be used easily and quickly in a booth while the interpreter is working is **Interplex UE**[5]. Instead of keeping isolated word lists, it allows to group all terms relating to a particular subject or field into multilingual glossaries that can be searched in an instant. This programme enables us to have several glossaries open at the same time, which is a very useful feature if the working domain is covered by more than one glossary. Similar to the previous analysed programmes, Interplex UE also allows us to import and export glossaries from and to Microsoft Word, Excel, and simple text files (see Fig. 3).

Interplex UE runs on Windows; nevertheless, it has a simpler version for iOS devices, one named Interplex Lite, for iPhone and iPod Touch, and another named Interplex HD, for iPad. Both glossaries and multi-glossary searchers offer the functionality of viewing expressions in each of the defined languages.

In general, Interplex UE has a user-friendly interface and it is regularly updated. It allows us to import and export glossaries from and to Microsoft Word and Excel formats. However, it, too, is platform dependent (Windows and iOS only), does not handle documents, only glossaries, and requires a commercial license.

The next two applications are particularly relevant for conference interpreting (simultaneous mode). **LookUp**[6] is a commercial multilingual glossary management tool developed for Windows, aiming to be used during simultaneous interpreting and while translating. It offers support for multilingual glossaries (English, German, Spanish, Italian and French), and its main purpose is to consult terminology rapidly while interpreting in a booth. **The Interpreter's Wizard**[7] is a free iPad application capable of managing bilingual glossaries in a booth. It is a simple, fast and easy-to-use application that helps the interpreter to search and visualise terminology in seconds.

Unit converters could also prove beneficial to interpreters when familiarising with new terminology measures such as temperature, distance, currency, acceleration, finance, speed, weight/mass and so on. **ConvertUnits**[8] and **OnlineConversion**[9] are two illustrative samples. Both seem to be quite comprehensive, providing online conversion calculators for all types of measurement units. Apart from this, interpreters can also find measure conversion tables for International System of Units, as well as calculators and conversers for units of acceleration, angles, area, energy, density force, power and pressure, astronomical units, clothing sizes, cooking volume units, mapping and navigation units, flowrates, etc. For Windows, there's **Convert**[10], and for Mac OS X, there's **Converto**[11]. These are two free and easy-to-use unit conversion programmes that convert the most popular units (additionally, Convert includes the ability to create custom conversions). There are also several mobile applications that can be also used, such as **Convert Units for Free**[12] and **Units**[13] for iOS devices, or **Unit Converter**[14] and **ConvertPad**[15] for Android devices.

Finally, **corpora** and corpus management tools (CMT) have proved most beneficial for interpreters as a device to speed up the preparation phase and to improve the quality of the input. A corpus can provide vast amounts of domain expert knowledge and accurate terminological and phraseological information in an efficient, effortless and inexpensive way.

3 Note-taking applications

Consecutive interpreters use a specific system of taking notes to retrieve part of their source speech understanding from memory while minimising the processing effort. This supporting technique is usually performed manually (pen and paper) and will continue in this manner for many years to come. However, as more and more interpreters are turning to mobile devices to take notes, it is just natural that those devices become the favourite note-taking and ubiquitous capture tool

[5] www.fourwillows.com
[6] www.lookup-web.de
[7] http://the-interpreters-wizard.topapp.net
[8] www.convertunits.com
[9] www.onlineconversion.com
[10] joshmadison.com/convert-for-windows
[11] http://fiplab.com
[12] www.freetheapps.com
[13] http://homegrownsw.com/units
[14] http://androidboy1.blogspot.com.es
[15] www.mathpad.com

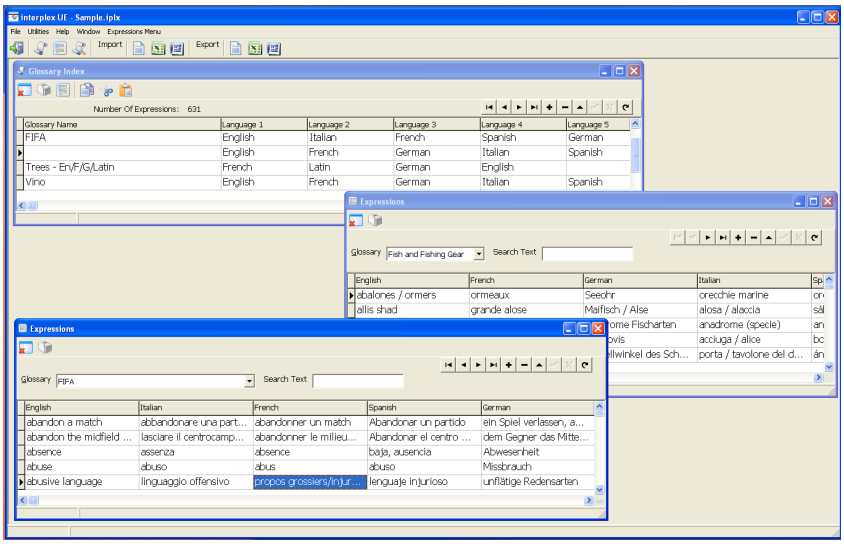


Figure 3: *Intraplex UE* screenshot.

on the go. In what follows, a good number of automated note-taking devices are presented.

Evernote[16] is a very dynamic and useful tool to keep more effective notes. It allows us to create an agenda note for each event, including any file, snapshot of handwritten note, audio message, Web page, PDF or Microsoft document. Evernote can also be used to work in a team, to keep event agendas in a shared business notebook so everyone can access the details of upcoming events, and to review action items that result from these events. With Evernote everything is shareable and accessible across all platforms. **Inkeness**[17] is also a very useful tool to write down ideas, take notes and make sketches. **Penultimate**[18] is similar, but, in addition, it allows the organisation of notes in notebooks. Inkeness and Punultimate are only available for iPad devices, and both enable sharing through Evernote and by e-mail. **LectureNotes**[19] and **PenSupremacy**[20] are two similar applications for Android. **My BIC Notes**[21] is an application specially designed for Android and iOS tablets. This application provides a set of tools for holding notes, drawing quick ideas or even doodles. In addition, it offers the functionality of adding sticky notes with personalised text, pictures and geometric shapes to the notes then printing them or sharing with others via e-mail.

Along the same line, there is a computer-assisted tool for semi-automation of the note-taking in consecutive interpreting that Rafajlovska (2013) discusses in her paper *Natural Language Processing Approach for Macedonian-French and Macedonian-English Interpreting based on Oral Sociopolitical Corpora*. This application provides a keyword with the most frequent symbols used by consecutive interpreters, which are linked to two *ad hoc* parallel dictionaries (Macedonian/English and Macedonian/French). By using the keyword, consecutive interpreters can take the same notes as they could on paper, but then they can also convert those notes into a readable message and save it for future reference.

Finally, digital pens appear to be the answer to the demand for dynamic technology capable of synchronising writing with ambient sound. Today these pens use real ink and write on real paper. **Sky Wifi Smartpen**, **Echo Smartpen** and **Livescribe** commercialised by Livescribe Inc.[22] and **Equil JOT**[23] are just some examples of smart digital pens. These four pens are capable of linking the written notes with ambient sound and uploading it to a computer over Bluetooth, Wireless or USB. Additionally, the provided

[16] https://evernote.com
[17] www.fenrir-inc.com
[18] http://evernote.com/penultimate
[19] www.acadoid.com
[20] https://sites.google.com/site/debarshishomepage
[21] www.bicworld.com
[22] www.livescribe.com
[23] www.myequil.com

software can be used to fully exploit the OCR capabilities of the pen and, for example, build glossaries. Another advantage of digital pens is the freedom to focus on listening and participating instead of worrying about catching every word during an event.

4 Voice recording and interpreter training

There are currently a number of applications that allow voice recording for training practice. Useful applications for managing text and audio files are **GoodReader**[24] and **Documents**[25]. Both tools allow the organisation, annotation and synchronisation of files of text (.TXT, .PDF), images, sound or video. They are available for iOS devices. Applications with a dual function are **Audacity**[26], **Adobe Audition**[27], **AudioNote**[28], **Notability**[29], **QuickVoice**[30], **Voice Dictation**[31], **Voice Pro**[32], amongst others. Besides voice recording, they allow the conversion into several audio formats, editing and quality improvement. Some of these tools provide interesting functionalities. For example, **AudioNote**[33], developed for multi-platforms (Windows, Mac OS X, Android and iOS), and **Notability**[34], for iOS, are interesting types of note-taking applications. Both are simple but powerful tools that combine the functionality of a notepad with voice recorder – a perfect choice for interpreters requiring a tool to synchronise text, drawings, photos, or handwritten notes with audio.

Simpler but equally useful, **Voice Dictation**[35], for iOS and **Voice Pro**[36] for Android, are two examples of easy-to-use voice recognition applications. Instead of typing, both applications use the microphone to convert audio notes to text automatically, which is very convenient to plan things to do, appointments and notes on the go.

Text-to-speech apps for iPad can also be successfully applied to teaching and improving language skills. For example, **Speak it!**[37], **Web Reader HD**[38], **Voice Dream Reader**[39], **Voxdox**[40] and **Talk - Text to Voice**[41] allow users to listen to words, texts, e-mail in several languages and formats. They are also available for Mac OS X, Windows, iOS and Android.

Finally, there is a very limited set of integrated tools that assist interpreters during their services or when training. **Black Box** (Sandrelli, 2005) is a computer-assisted interpreter training tool designed to help interpreters work with a range of different materials (texts, audio, video, different types of exercises) and store their results for later review. It can be used to practice in different ways: either by interpreting some audio or video clips or by doing some practical interpreting exercises, such as shadowing, cloze exercises or sight translation. It also allows teachers to edit and break down video and audio recordings to create different exercises and adapt authentic conference materials to the students' level of expertise. Black Box can be considered a suitable training workbench for trainee interpreters.

Other web-based environments have recently been created along similar lines. **InterpretaWeb**[42] and **Linkinterpreting**[43] provide interpreters and students with a wide range of exercises (cloze, memory, cluster), and complete speeches to practice simultaneous and consecutive interpreting, along with information resources and news related to interpreting. These websites are of great use to students and for novice interpreters who are willing to practice and improve their interpreting skills.

[24] www.goodiware.com
[25] http://readdle.com
[26] http://audacity.sourceforge.net
[27] www.adobe.com/products/audition.html
[28] http://luminantsoftware.com/iphone/audionote.html
[29] www.gingerlabs.com
[30] www.nfinityinc.com/quickvoiceip.html
[31] https://itunes.apple.com/us/app/voice-dictation-voice-to-sms/id492594590?mt=8
[32] www.voicepro.it
[33] http://luminantsoftware.com
[34] www.gingerlabs.com
[35] http://quanticapps.com
[36] http://forum.voicepro.it
[37] https://itunes.apple.com/us/app/speak-it!-text-to-speech/id308629295?mt=8
[38] https://itunes.apple.com/us/app/web-reader-hd-text-to-speech/id376528713?mt=8
[39] www.voicedream.com
[40] www.voxdox.net
[41] https://plus.google.com/communities/107986392540899459664
[42] www.interpretaweb.es
[43] http://linkterpreting.uvigo.es

5 Conclusion

Technology tools open up a new world of possibilities for interpreters. This paper has presented an overview of tools and applications available for interpreting practice and training. Although the number of these technologies is growing fast due to an increasing interest towards interpreters' needs, they are still insufficient and unable to fulfil all the necessary requirements. There is an urgent need to develop technologies that automate the process, increase the productivity and ease the labour-intensive activities of an interpreter (either in the preparation stage, before their interpreting service or during it). A next step in the right direction could be to gather detailed information to better ascertain interpreters' technology awareness and real needs in order to design new tools and improve existing ones.

References

Rafajlovska, A. (2013). *Natural Language Processing Approach for Macedonian-French and Macedonian-English Interpreting based on Oral Sociopolitical Corpora*. Master Thesis, Université de Franche-Comté, France and Universidade do Algarve, Portugal.

Roberts, R. (1994). Community Interpreting Today and Tomorrow. In Krawutschke, P., editor, *35th Annual Conf. of the American Translators Association*, pages 127–138. Medford, NJ: Learned Information.

Costa et al. (2014a)

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014). **A comparative User Evaluation of Terminology Management Tools for Interpreters**. In *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, pages 68-76, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

A comparative User Evaluation of Terminology Management Tools for Interpreters

Hernani Costa* **Gloria Corpas Pastor** **Isabel Durán Muñoz**
LEXYTRAD, University of Malaga, Spain
{hercos,gcorpas,iduran}@uma.es

Abstract

When facing new fields, interpreters need to perform extensive searches for specialised knowledge and terminology. They require this information prior to an interpretation and have it accessible during the interpreting service. Fortunately, there are currently several terminology management tools capable of assisting interpreters before and during an interpretation service. Although these tools appear to be quite similar, they provide different kind of features and as a result they exhibit different degrees of usefulness. This paper aims at describing current terminology management tools with a view to establishing a set of features to assess the extent to which terminology tools meet the specific needs of the interpreters. Subsequently, a comparative analysis is performed to evaluate these tools based on the list of features previously identified.

1 Introduction

Professional interpreters frequently face different settings and specialised fields in their interpretation services and yet they always need to provide excellent results. They might be called to work for specialists that share a background knowledge that is totally or partially unknown to laypersons and/or outsiders (Will, 2007). When interpreters lack the necessary background knowledge or experience, they usually need to perform extensive searches for specialised knowledge and terminology in a very efficient way in order to supply this deficit and acquire the required information.

Even though there are several modes of interpretation, depending mainly on the timing/delay of the interpretation, the direction and the setting (cf. Pöchhacker, 2007), it is not possible for interpreters to collect the relevant specialised information during the interpretation service itself. Interpreters are required to find the necessary information prior to interpretation and have it accessible during the service, even though they sometimes are able to carry out searches during the service.

According to Rodríguez and Schnell (2009), terminology work is present in the whole process of preparation prior to an interpretation service. For example, interpreters become familiar with the subject field by searching for specialised documents, by extracting terms and looking for synonyms and hyperonyms, by finding and developing acronyms and abbreviations and by compiling a glossary. According to these authors, interpreters tend to compile in-house glossaries tailor to their individual needs as the main way to prepare the terminology of a given interpretation.

2 Interpreters' Needs

The potentialities of computers for improving interpreters' working conditions was realised a long time ago by Gile (1987). However, very little progress has been made so far. Costa, Corpas Pastor and Durán-Muñoz (2014) offer a tentative catalogue of current language technologies for interpreters, divided into terminology tools for interpreters, note-taking applications for consecutive interpreting, applications for voice recording and training tools. This paper focus exclusively on terminology tools for interpreters with a view to performing a user evaluation.

As a rule, most interpreters seem to be unaware of the opportunities offered by language technologies. As far as terminology is concerned, interpreters continue to store information and terminology on scraps

*Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement N° 317471.

of paper or excel spreadsheets, while the use of technologies and terminology management tools is still very low. A study conducted by Moser-Mercer (1992:507, quoted in Bilgen, 2009) rejected the assumption that "interpreters' needs are identical to those of translators and terminologists" and intended to "survey how conference interpreters handle terminology documentation and document control and to offer some guidelines as to the interpretation-specific software tools for terminology and documentation management". The results of this study includes some key findings, such as the conclusion that most of the respondents were interested in exchanging terminological information and that they were open to using computers in their profession. According to these findings, Moser-Mercer (1992) highlighted that "software developers targeting the conference interpreting market must provide a tool that meets the specific needs of the interpreters and not just market translation tools" (ibid:511). More recent studies have also studied interpreters' current needs and practices regarding terminology management (Rodríguez and Schnell, 2009; Bilgen, 2009), and they also share the same findings: interpreters require specific tools to meet their needs, which are different from translators and terminologists. According to a survey conducted by Bilgen (2009), 85% of respondents are open to using computers, yet conventional methods still prevail over the use of computerised methods of terminology management. The author observed that respondents had no or little experience with terminology management software, and those with some experience were most dissatisfied with the money and time they had to invest in them, and their overall experience was mediocre (ibid:66). Respondents indicated that their priorities were different from those identified in terminology literature in terms of terminological information stored, and the way in which term records are structured. This is an important aspect that differentiates the needs of interpreters and translators as regards definitions and contexts (Bilgen, 2009). Due to their working conditions, translators usually prefer to consult multiple definitions and contexts to find the best solution for the translation problem. On the contrary, interpreters will rarely have the time to go over multiple definitions, contexts, etc. to find the right one, and thus, they will need to store the most concise information to be able to consult it in the quickest and easiest way. Their responses in this survey also showed that the way they retrieve terminological information was context-specific, and that there was also a significant variation among individual interpreters. Flexibility is, therefore, of great importance to interpreters due to the variation of their context-specific terminology management practices, and on their individual preferences regarding the storage, organisation and retrieval of terminological information (ibid:92). Rodríguez and Schnell (2009), after a thorough analysis of interpreters' needs and in order to meet their requirements as regards terminology management tools, propose the possibility of developing small databases that vary according to the area of speciality or according to the conference and client. These mini-databases would be multilingual and include an option allowing the interpreter to switch the source and target languages. This assumption is in line with the Function Theory (Bergenholtz and Tarp, 2003; Tarp, 2008) and electronic multifunctional dictionaries (Spohr, 2009), which both defend the need to elaborate terminological entries according to the potential users. Rodríguez and Schnell (2009) recognise five features that would distinguish the interpreters' mini-databases from the terminology databases intended for translators: speed of consultation; intuitive navigation; possibility of updating the terminology record in the interpretation booth; considerable freedom to define the basic structure; and multiple ways of filtering data.

Accordingly, they also suggest the abandonment of the usual terminology methodology if the intention is to provide interpreters with specific glossaries tailored to their needs. The authors advance the use of a semasiological and associative methodology instead of the onomasiological approach, as "it does not adapt well to interpretation because the cognitive effort required by the onomasiological structures slows down the interpretation process" (ibid).

3 Terminology Management Tools for Interpreters

There are some specialised computer and mobile software that can be used to quickly compile, store, manage and search within glossaries. The most outstanding applications developed by/for interpreters are described in detail below. They can be typically used to prepare an interpretation, in consecutive interpreting or in a booth. These applications are quite similar to the look-up terminology tools currently

used by translators (Durán Muñoz, 2012). In fact, some of them have been developed to cater for the needs of both translators and interpreters. Due to the lack of space, this article is focused on standalone applications, but other types of applications like Web-based (e.g. Interpreters' Help[1]) can also be used for the same purposes (Ruetten, 2014).

Intragloss[2] is a commercial Mac OS X software created specifically to help interpreters when preparing for an event by allowing them to manage glossaries. This application can be simply defined as a glossary and document management tool created to help the interpreter prepare, use and merge different glossaries with preparation documents, in more than 180 different languages. It allows to import and export glossaries from and to plain text, Microsoft Word and Excel formats. Every glossary imported to, or created in, is assigned to a domain glossary (considered the highest level of knowledge), which contains all the glossaries from the sub-areas of knowledge, named 'assignments'. The creation of an assignment glossary can be done in two different ways: either by extracting automatically all the terms from the domain glossary that appear in the imported documents, or by highlighting a term in the document, search for it on search sites (such as online glossaries, terminology databases, dictionaries and general Web pages) and manually add the new translated term to the assignment glossary. It is important to mention that the online search can be made within Intragloss. Another interesting feature is that Intraglosss permits to copy assignment glossaries and assignment entries from one assignment to another. The domain glossary may be multilingual as it can include several bilingual assignment glossaries. By a way of example, if we have two assignment glossaries English/French and Dutch/English, in the same domain, the domain glossary will be French/English/Dutch, i.e. multilingual. Finally, Intragloss also allows to manually add meta-information to each glossary entry (see Fig. 1a).

In short, Intragloss is an intuitive and easy-to-use tool that facilitates the interpreters' terminology management process by producing glossaries (imported or created ad hoc), by searching on several websites simultaneously and by highlighting all the terms in the documents that appear in the domain glossary. However, it is currently platform dependent and only works on Mac OS X platforms.

InterpretBank[3] is a simple terminology and knowledge management software tool designed both for interpreters and translators using Windows and Android. It helps to manage, learn and look up glossaries and term-related information. Due to its modular architecture (see Fig. 1b), it can be used to guide the interpreter during the entire workflow process, starting from the creation and management of multilingual glossaries (TermMode), passing through the study of these glossaries (MemoryMode), and finally allowing the interpreter to look up terms while in a booth (ConferenceMode). InterpretBank also has an Android version called InterpretBank Lite. This application is specifically designed to access bi- or trilingual glossaries previously created with the desktop version. It is useful when working as a consecutive, community or liaison interpreter, when a quick look up at the terminology list is necessary.

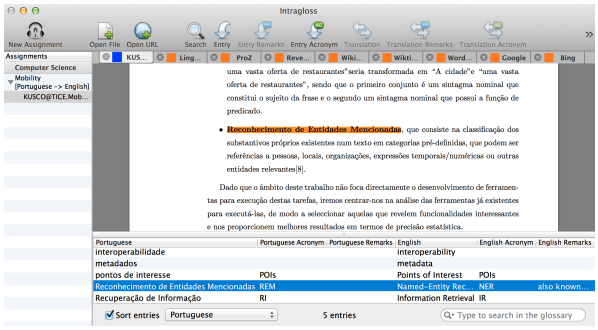
InterpretBank has a user-friendly, intuitive and easy-to-use interface. It allows us to import and export glossaries in different formats (Microsoft Word, Microsoft Excel, simple text files, Android and TMEX) and suggests translation candidates by taking advantage of online translation portal services, such as Wikipedia, MyMemory and Bing. However, it is platform-dependent (it only works on Windows and Android), does not handle documents, only glossaries and requires a commercial license.

Interplex UE[4] is a user-friendly multilingual glossary management programme that can be used easily and quickly in a booth while the interpreter is working. Instead of keeping isolated word lists, it allows to group all terms relating to a particular subject or field into multilingual glossaries that can be searched in an instant. As we can see in Fig. 1c, this programme permits to have several glossaries open at the same time, which is a very useful feature if the working domain is covered by more than one glossary. Similar to the previous analysed programmes, Interplex UE also allows to import and export glossaries from and to Microsoft Word, Excel, and simple text files. Interplex UE runs on Windows; nevertheless, it has a simpler version for iOS devices, one named Interplex Lite, for iPhone and iPod Touch, and another named Interplex HD, for iPad. Both glossaries and multi-glossary searchers offer the functionality of

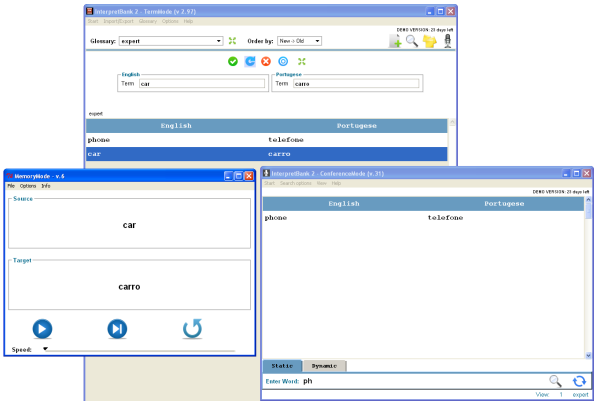
[1] www.interpretershelp.com
[2] intragloss.com
[3] www.interpretbank.de
[4] www.fourwillows.com

viewing expressions in each of the defined languages.

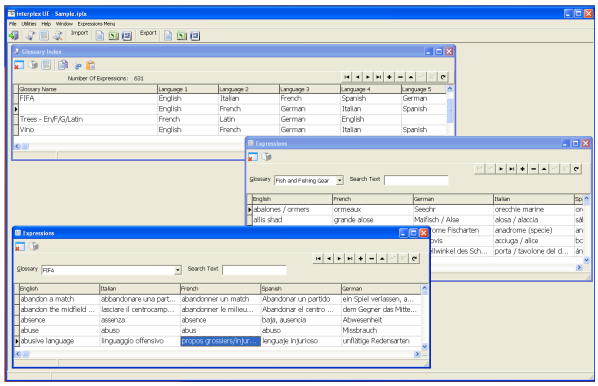
In general, Interplex UE has a user-friendly interface and it is regularly updated. It allows to import and export glossaries from and to Microsoft Word and Excel formats. However, it, too, is platform dependent (only works on Windows and iOS), does not handle documents, only glossaries, and requires a commercial license.



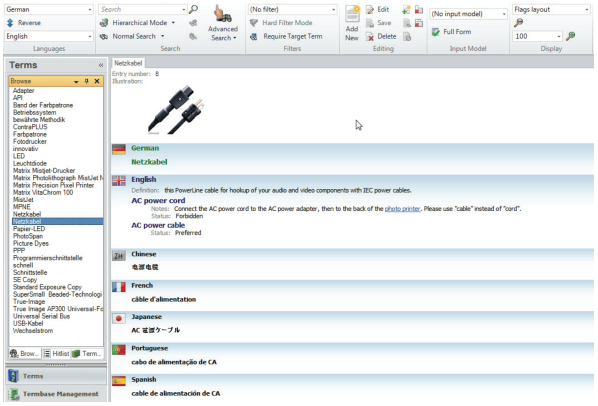
(a) Intragloss.



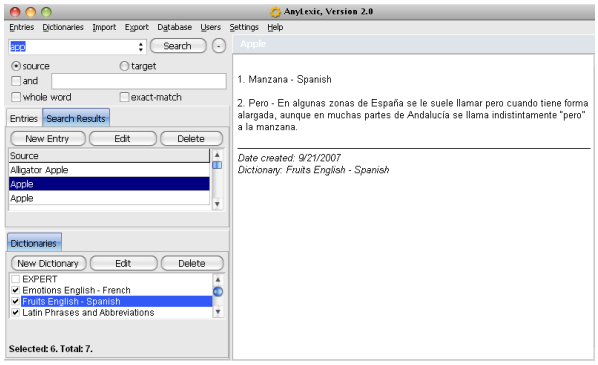
(b) InterpretBank.



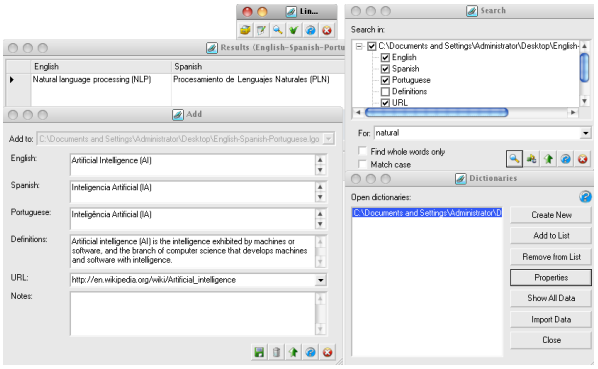
(c) Intraplex.



(d) SDL MultiTerm.



(e) AnyLexic.



(f) Lingo.

Figure 1: Screenshots of various terminology managment tools.

SDL MultiTerm Desktop[5] is a commercial terminology management tool developed for Windows that provides one solution to store and manage multilingual terminology. MultiTerm was first launched in 1990 by Trados GmbH but in 2005 the company was acquired by SDL[6], which renamed MultiTerm to SDL MultiTerm. Today, SDL MultiTerm is a terminology management tool commercialised by SDL as a standalone application, which has been improved according to the translators' needs. Alternatively, MultiTerm can be used within the SDL Trados Studio[7] as an integrated tool. As translators can

[5] www.sdl.com/products/sdl-multiterm/desktop.html
[6] www.sdl.com
[7] www.sdl.com/products/sdl-trados-studio

easily edit and add terminology within SDL Trados Studio, MultiTerm helps to improve the efficiency of the translation process and promotes high-quality translated content with real-time verification of multilingual terminology. This application is very complete because it allows to store an unlimited number of terms in a vast number of languages; imports and exports glossaries from and to different technology environments, such as Microsoft Excel, XML, TBX and several other proprietary formats; permits to manually add a variety of meta-data information, such as synonyms, context, definitions, associated project, part-of-speech tags, URLs, etc. Apart from the previous mentioned descriptive fields, MultiTerm also allows the user to insert illustrations to the terms in the terminology database (which can be stored either locally or, for collaborative purposes, in a remote server). It is important to mention that this visual reference feature is very useful specially to interpreters and translators dealing with unfamiliar terms. Moreover, MultiTerm has an advanced search feature that permits to search not only the indexed terms but also in their descriptive fields, or create filters to make custom searches within specific fields, like language, definition, part-of-speech, etc. Nevertheless, the most interesting feature about MultiTerm is its concept-oriented feature, i.e. each entry in MultiTerm corresponds to a single concept, which can be described by different terms in both source and target language. This detail is very important because it allows the user to centralise and customise the terms with more information, such as different possible translations and their corresponding contexts (see Fig. 1d).

In general, MultiTerm can be seen as an advanced multilingual terminology tool with an intuitive and easy-to-use interface. Although MultiTerm was originally designed for translators, it can also be used by interpreters. Its main advantage to interpreters, when compared with other terminology tools, is twofold: it allows to add several translation terms in one entry and permits to customise a wide variety of descriptive fields, such as illustrations, associated projects, definitions, etc. However, it can only be used on Windows, does not handle documents and there is no demo version available.

AnyLexic[8] is an easy-to-use terminology management tool developed for Windows with a simple and intuitive interface. It was not designed to tie any particular terminological requirement, instead it aims to help the interpreter prepare, use and manage different glossaries or dictionaries. AnyLexic can be described as a robust terminology management tool, as it enables users to easily create and manage multiple mono-, bi- or multilingual glossaries in any language and to import and export glossaries from and to Microsoft Excel, plain text and AnyLexic Exchange Format (AEF). In addition, each entry in the glossary can have multiple translation equivalents in the target language along with notes. The search for records in the database allow users to combine different options, such as search for all source terms or translation candidates and associated notes. In addition, the search can be performed within one or multiple glossaries (see Fig. 1e). Another interesting feature in AnyLexic is the way that records can be displayed using different templates with configurable text colour, background colour, font size and text format. Besides, it is possible to create our own template for displaying the records. With the purpose of simplifying the teamwork process, this tool has an additional option to exchange any glossary with other AnyLexic users by either using the AEF proprietary format or by accessing a remote glossary, a very useful feature when co-operating with other interpreters or translators on a project.

In general, AnyLexic is an easy and convenient terminology database managing software for working with terminology, creating, editing and exchanging glossaries when working under one project both alone or with other working partners. However, it only works on Windows platforms and even though an evaluation version is available for 30 days, it requires a commercial license.

Lingo[9] is a commercial Windows terminology management tool designed to create and manage terminology databases, whether mono- or multilingual. It can import from and export to TMX and plain text. Its main features are: creation and management of any number of specialised glossaries/dictionaries in any language; can handle large files (i.e. over 50K entries); it allows users to have several glossaries open at the same time; and it has a rapid and easily configurable search functionality that can be customised to search for all terms, translation candidates and associated descriptive fields, either in all glossaries or in a specific one. Another interesting feature is the drag and drop functionality, which

[8] www.anylexic.com

[9] www.lexicool.com/soft_lingo2.asp

allows to easily insert words into Microsoft documents, for instance.

As we can see in Fig. 1f, Lingo is a simple and user-friendly software that offers an effective way to create and manage multilingual glossaries in any language. Additionally, it permits to manually add an infinite number of customised fields into each entry, such as definitions, URLs, synonyms, antonyms, contextual information, notes or any other desirable field. However, it is platform dependent and does not import from or export to common formats like Microsoft Word or Excel.

UniLex[10] is a free terminology management tool created by Acolada GmbH for Windows. It aims to help interpreters and translators prepare, use and manage bilingual glossaries or dictionaries in approximately 30 different languages. UniLex offers a variety of search functions and the possibility to combine user glossaries or dictionaries with a full range of dictionaries available in the UniLex series (e.g. Blaha: Pocket Dictionary of Automobile Technology German/English), which can be acquired as single user versions or as network versions for collaborative purposes. UniLex can also be used in a network environment, which allows users to exchange glossaries or dictionaries. Nevertheless, this additional feature requires a commercial license.

In general, UniLex is not only capable of managing user bilingual glossaries or dictionaries, but also dictionary titles from renowned publishers, which are sold by the company to be consulted within UniLex. However, it only works on Windows and does not handle multilingual glossaries.

The Interpreter's Wizard[11] is a free iPad application capable of managing bilingual glossaries in a booth. It is a simple, fast and easy-to-use application that helps the interpreter to search and visualise terminology in seconds. The system includes rapid and easily configurable search functionality that can be customised to search for all terms, translation candidates either in all glossaries or in a specific one. Nevertheless, all the imported glossaries need to be previously created and converted online to the proprietary format, and it does not allow users to export glossaries.

3.1 Comparative Analysis

Despite the aforementioned terminology tools can be used to prepare a given interpretation according to the interpreters' requirements identified in section 2, these systems differ from one another in their functionalities, practical issues and degrees of user-friendliness. Therefore, it is necessary to establish a set of specific and measurable features that permit us to assess and distinguish the different tools concerning users' needs in such a way that the results would be useful for both potential users as well as to the designers of such systems. Departing from the conclusions drawn from the literature review (see section 2) and the description of the terminology tools analysed in section 3, we provide in this section an analysis of these tools based on our own practical set of measurable features. For instance, the "freedom to define the basic structure" identified by Rodríguez and Schnell (2009) was reformulated into several practical measurable features, such as "Nº of descriptive fields", "Nº of working languages" and "Nº of languages per glossary". Moreover, the possibility of "developing multilingual mini-databases", also identified in their study, was reconsidered as measurable features by means of the following criteria: "Manages multiple glossaries" and "Nº of languages per glossary". Another example is the "Remote Glossary Exchange" measurable feature, which was inferred from the study conducted by Bilgen (2009), who identified the need to exchange terminological information.

After a careful analysis of the priorities for the design and features to be included in a terminology management tool for interpreters, 15 features were identified, 5 of which were classified as fundamental to a terminology tool (10 points) and 10 as secondary (5 points). Then, these features were used to evaluate the eight tools presented in section 3 and to investigate which one is the most complete. The first considered feature clarifies if the tools were designed to handle multiple glossaries in their interfaces at same time (**Manages multiple glossaries**). The next two features are somehow related. The **Nº of possible working languages** describes how many different working languages are permitted by the application. Then, considering these working languages, how many of them can be used at the same time per glossary (**Nº of languages per glossary allowed**). The next feature is related with

[10] www.acolada.de/unilex.htm

[11] the-interpreters-wizard.topapp.net

Feature/ Evaluation Criteria	**Intragloss** Pre-1.0 (2014)	**InterpretBank** 3.102 (2014)	**Intraplex** 2.1.1.47 (2012)	**SDL MultiTerm** 2014 (2013)	**AnyLexic** 2.0.0.2110 (2009)	**Lingo** 4 (2011)	**UniLex** 0.9 (2007)	**The Interpreter's Wizard** 2.0 (2011)
Manages multiple glossaries ⟨no=0; yes=10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	no ⟨0⟩	yes ⟨10⟩
Nº of possible working languages (≤100=4; >100=7; unlimited=10⟩	≈180 ⟨7⟩	≈35 ⟨4⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	≈30 ⟨4⟩	unlimited ⟨10⟩
Nº of languages per glossary allowed ⟨≤3=5; unlimited=10⟩	2 ⟨5⟩	2 ⟨5⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	2 ⟨5⟩	2 ⟨5⟩
Nº of descriptive fields ⟨non=0; 1=3; [2-5]=7; >5=10⟩	4 ⟨7⟩	4 ⟨7⟩	non ⟨0⟩	>5 ⟨10⟩	1 ⟨3⟩	>5 ⟨10⟩	2 ⟨7⟩	non ⟨0⟩
Handles documents ⟨no=0; yes=10⟩	yes (PDF, MS Word, Pages and Keynote files) ⟨10⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩
Unicode compatibility ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩	yes ⟨5⟩
Imports from ⟨1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	MS Word, Excel and Plain Text ⟨3⟩	MS Word, Excel, TMEX and Plain Text ⟨4⟩	MS Word, Excel and Plain Text ⟨3⟩	MS Word, Excel and other CAT formats ⟨5⟩	MS Excel, Plain Text and AEF ⟨3⟩	TMX and Pain Text ⟨2⟩	Plain Text ⟨1⟩	Proprietary format ⟨1⟩
Exports to ⟨non=0; 1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	MS Word and Excel ⟨2⟩	MS Word, Excel, TMEX, Android and Plain Text ⟨4⟩	MS Word, Excel and Plain Text ⟨3⟩	MS Word, Excel and other CAT formats ⟨5⟩	MS Excel, Plain Text and AEF ⟨3⟩	TMX and Plain Text ⟨2⟩	Plain Text ⟨1⟩	non ⟨0⟩
Embedded online search for translation candidates ⟨>no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩
Interface's supported languages ⟨1=1; [2-5]=3; >5=5⟩	English ⟨1⟩	English ⟨1⟩	English ⟨1⟩	English, German, French, Spanish, Japanese and Simplified Chinese ⟨5⟩	English, Simplified Chinese, German, Spanish, French, Italian, Dutch, Polish, Portuguese, Russian and Ukrainian ⟨5⟩	English ⟨1⟩	English, German, French and Spanish ⟨3⟩	English ⟨1⟩
Remote Glossary Exchange ⟨no=0; yes=5⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩
Well-documented ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩	no ⟨0⟩
Availability ⟨proprietary without demo=1; proprietary with demo=3; free=5⟩	proprietary with demo ⟨3⟩	proprietary with demo ⟨3⟩	proprietary with demo ⟨3⟩	proprietary without demo⟨1⟩	proprietary with demo ⟨3⟩	proprietary with demo ⟨3⟩	free ⟨5⟩	free ⟨5⟩
Operating System(s) ⟨1=1; 2=3; ≥3=5⟩	Mac OS X ⟨1⟩	Windows and Android ⟨3⟩	Windowns and iOS ⟨3⟩	Windows ⟨1⟩	Windows ⟨1⟩	Windows ⟨1⟩	Windows ⟨1⟩	iOS (only iPad) ⟨1⟩
Other relevant features ⟨subjective analysis=max. 5⟩	allows to highlight terms in the documents and merge a glossary with a document making it annotated to be printed ⟨5⟩	the MemoryMode helps to memorise bilingual glossaries ⟨4⟩	permits to have several glossaries open at the same time ⟨2⟩	it is a concept oriented-tool and permits to add illustrations into each entry ⟨5⟩	allows to share glossaries within a group of AnyLexic users ⟨1⟩	permits to add an unlimited number of descriptive fields ⟨2⟩	–	quick performance ⟨1⟩
Final Mark	69	60	55	77	64	61	27	39

Table 1: Comparative view and classification of several terminology management tools.

all types of descriptive fields that these tools allow to add to each glossary entry (**N⁰ of descriptive fields**). The possibility of managing terminology with preparation documents (**Handles documents**) is another relevant feature for interpreters seeking for tools capable of highlighting terms in documents, for example. Equally import is the Unicode support (**Unicode compatibility**) as it provides a unique number for every character, no matter what the platform, the program, or the language is. In other words, an application that supports full Unicode means that it has support for any ASCII or non-ASCII language, such as Hebrew or Russian, two non-ASCII languages. **Imports from** and **Exports to**, as its name suggests, represents the supported input and output formats. The **Embedded online search for translation candidates** is a relevant add-in for terminology tools, as it permits to focus the search for terminological candidates within the tool. Despite the fact that all the tools have English as a default language, the support of multiple languages (**Interface's supported languages**) is another important feature as it allows to increase the number of potential users that a terminology tool can reach. The **Remote glossary exchange** feature is important when co-operating with other working partners remotely is required. The next three features are related with the available documentation, their availability and platform dependency (**Well-documented**, **Availability** and **Operating System(s)**, respectively). Finally, the last row presents some unique characteristics along with some relevant comments (**Other relevant features**).

Based on this comparative analysis, none of the investigated terminology tools exhibit all the proposed features. Nevertheless, SDL MultiTerm and Intragloss are the best classified with 77 and 69 points out of 100, respectively. This is not surprising because SDL MultiTerm is the most expensive tool nowadays available on the market and, apart from that, it has been developed for more than 20 years. Its flexibility to easily store, manage and search for multilingual terminology and definitions is just an example of the features that meet the specific needs of an interpreter. The score of Intragloss, released last year as a beta version, is neither surprising due to its novelty and design purposes, i.e. it was developed by interpreters for interpreters and thus corresponds better to their needs. On the other hand, UniLex and The Interpreter's Wizard tools got the worst scores due to the lack of features offered. About the remaining tools (AnyLexic, Lingo, InterpretBank and Intraplex) we can say that they have similar features, which resulted in similar scores (64, 61, 60 and 55, respectively).

4 Conclusion

This paper presents an overview of the most relevant features that terminology management tools should have in order to help interpreters before and during the interpretation process. Eight terminology tools are discussed and a comparative analysis is performed to evaluate them on the bases of the set of features previously identified. This comparative analysis not only aims at highlighting some of the features that interpreters can expect from the currently available terminology management tools on the market, but also intended to help them choose a specific tool for a given interpretation project. Table 1 provides interpreters with a comprehensive and up-to-date review of the currently available terminology tools on the market. It is envisaged to serve as a concise guide to help interpreters choose the terminology management tool that best caters for their specific needs, in order to help them work more efficiently, store and share terminology more easily, as well as save time when a looking for a specific feature most suited to a specific interpreting service.

Although most of the analysed tools could be considered to be very flexible when searching for terminology within glossaries, it appears that none of them fulfil all needs of interpreters. It is worth mentioning that some tools require a steep learning curve while others imply a significant financial investment (e.g. Lingo and SDL MultiTerm, respectively). Moreover, some tools are fairly basic and more orientated towards creating and managing bilingual or multilingual glossaries rather than more comprehensive terminology records with supporting information (e.g. UniLex and The Interpreter's Wizard).

Given that quality terminology management ranks high in their priorities, it would seem that there is a pressing need to design terminology management tools tailored to assist interpreters in the preparation stage, before their interpreting service or during it. In this respect, it would be necessary to identify the exact needs of interpreters (which are likely to be different from translators).

Acknowledgements

The research reported in this article has been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (nº FFI2012-38881, 2012-2015), and the R&D Project for Excelence TERMITUR (ref. nº HUM2754, 2014-2017).

References

Henning Bergenholtz and Sven Tarp. 2003. Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. *Hermes, Journal of Linguistics*, 31:171–196.

Baris Bilgen. 2009. *Investigating Terminology Management for Conference Interpreters*. MA dissertation, University of Otawa, Otawa, Canada.

Hernani Costa, Gloria Corpas Pastor, and Isabel Durán Muñoz. 2014. Technology-assisted Interpreting. *MultiLingual 143*, 25(3):27–32, April/May.

Isabel Durán Muñoz. 2012. Meeting Translators' Needs: Translation-oriented Terminological Management and Applications. *The Journal of Specialised Translation*, 18:77–92. Available at: `http://www.jostrans.org/issue18/art_duran.pdf` (Accessed 30 June 2014).

Daniel Gile. 1987. La terminotique en interprétation de conférence: un potentiel à exploiter. *Meta: Translators' Journal*, 32(2):164–169, June.

Barbara Moser-Mercer. 1992. Banking on Terminology: Conference Interpreters in the Electronic Age. *Meta: Translators' Journal*, 37(3):507–522, September.

Franz Pöchhacker. 2007. *Introducing Interpreting Studies*. London and New York: Routledge, 2^{nd} edition.

Nadia Rodríguez and Bettina Schnell. 2009. A Look at Terminology Adapted to the Requirements of Interpretation. *Language Update*, 6(1):21–27. Available at: `http://www.btb.termiumplus.gc.ca/tpv2guides/guides/favart/index-eng.html?lang=eng&lettr=indx_autr8gijKBACeGnI&page=9oHAHvmFzkgE.html` (Accessed 30 June 2014).

Anja Ruetten. 2014, June 30. Booth-friendly terminology management revisited - 2 newcomers. Retrieved from: `http://blog.sprachmanagement.net/?p=305` (Accessed 30 June 2014).

Dennis Spohr. 2009. Towards a Multifunctional Electronic Dictionary Using a Metamodel of User Needs. In *eLexicography in the 21^{st} century: New challenges, new applications*, Louvain-La-Neuve, Belgium. Presses Universitaires de Louvain.

Sven Tarp. 2008. *Lexicography in the Borderland Between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Lexicographica: Series maior. Walter de Gruyter, 1^{st} edition.

Martin Will. 2007. Terminology Work for Simultaneous Interpreters in LSP Conferences: Model and Method. In Heidrun Gerzymisch-Arbogast and Gerhard Budin, editors, *Proc. of the Marie Curie Euroconferences MuTra: LSP Translation Scenario*, EU-High-Level Scientific Conference Series, Vienna, Austria.

Costa et al. (2014c)

Costa, H., Corpas Pastor, G., and Seghiri, M. (2014). **iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora**. In *Translating and the Computer 36 - AsLing*, London, UK.

iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora

Hernani Costa, Gloria Corpas Pastor and Miriam Seghiri
LEXYTRAD, University of Malaga, Spain
{hercos,gcorpas,seghiri}@uma.es

Abstract

This article presents an ongoing project that aims to design and develop a robust and agile web-based application capable of semi-automatically compiling monolingual and multilingual comparable corpora, which we named iCompileCorpora. The dimensions that comprise iCompileCorpora can be represented in a layered model comprising a manual, a semi-automatic and a Cross-Language Information Retrieval (CLIR) layer. This design option will not only permit to increase the flexibility of the compilation process, but also to hierarchically extend the manual layer features to the semi-automatic web-based layer and then to the semi-automatic CLIR layer. The manual layer presents the option of compiling monolingual or multilingual corpora. It will allow the manual upload of documents from a local or remote directory onto the platform. The second layer will permit the exploitation of either monolingual or multilingual corpora mined from the Internet. As nowadays there is an increasing demand for systems that can somehow cross the language boundaries by retrieving information of various languages with just one query, the third layer aims to answer this demand by taking advantage of CLIR techniques to find relevant information written in a language different from the one semi-automatically retrieved by the methodology used in the previous layer.

Keywords: Comparable Corpora, Corpora, Corpora Compilation Tools, Cross-Language Information Retrieval, Translation Technologies.

1 Introduction

The interest in mono-, bi- and multilingual corpora is vital in many research areas such as language learning, stylistics, sociolinguistics, translation studies, amongst other research areas. Particularly in translation, their benefits have been demonstrated by various authors (cf. Bowker and Pearson, 2002; Bowker, 2002; Zanettin et al., 2003; Corpas Pastor and Seghiri, 2009). The main advantages of its usage are their objectivity, reusability, multiplicity and applicability of uses, easy handling and quick access to large volume of data. In detail, corpus linguistics:

- Empowers the study of the foreign language: the study of the foreign language with the use of corpora allows the foreign language learners to get a better "feeling" about that language and learn the language through "real world" texts rather than "controlled" texts (cf. Gries, 2008).
- Simplifies the study of naturalistic linguistic information: as previously mentioned, a corpus assembles "real world" text, mostly a product of real life situations, which results in a valuable research source for dialectology (cf. Hollmann and Siewierska, 2006), sociolinguistics (cf. Baker, 2010) and stylistics (cf. Wynne, 2006), for example.
- Helps linguistic research: as the time needed to find particular words or phrases has been dramatically reduced with the use of electronically readable corpora, a research that would take days or even weeks to be manually performed can be done in a couple of seconds with an high degree of accuracy.
- Enables the study of wider patterns and collocation of words: before the advent of computers, corpus linguistics was studying only single words and their frequency. More recently, the emergence of modern technology allowed the study of wider patterns and collocation of words (cf. Roland et al., 2007).
- Allows simultaneous analysis of multiple parameters: in the last decades, the development of corpus linguistic software tools helped the researchers to analyse a wider number of parameters simultaneously, such as determine how the usage of a particular word and its syntactic function varies.

Moreover, they are a suitable tool for translators, as they can easily determine how specific words and their synonyms collocate and vary in practical use or even help interpreters speeding up the research for unfamiliar terminology (cf. Costa et al., 2014). Furthermore, in the last decade, a growing interest in bi- and multilingual corpora has been shown by researchers working in other fields, such as terminology and specialised language, automatic and assisted translation, language teaching, Natural Language Processing, amongst others. Nevertheless, the lack of sufficient/up-to-date parallel corpora and linguistic resources for narrow domains and poorly-resourced languages is currently one of the major obstacles to further advancement on these areas. One potential solution to the insufficient parallel corpora is the exploitation of non-parallel bi- and multilingual text resources, also known as comparable corpora (i.e. corpora that include similar types of original texts in one or more language using the same design criteria, cf. EAGLES, 1996; Corpas Pastor, 2001:158).

Even though comparable corpora can compensate for the shortage of linguistic resources and ultimately improve automated translations quality for under-resourced languages and narrow domains for example, the problem of data collection presupposes a significant technical challenge. Moreover, the difficulty of retrieving and classifying such data is considered a complex issue as there is no unique notion of what it really covers and how it can be truly exploited (cf. Skadina et al., 2010:12).

2 Existing Corpora Compilation Solutions

Although this compilation process could be manually performed, nowadays specialised tools can be used to automate this tedious task. By a way of example, BootCaT[1] (Baroni and Bernardini, 2004) was built to exploit specialised monolingual corpora from the Web. It is capable of compiling a corpus through automated search queries, and only requires a small set of seed words as input. This tool has been used, for example, to create specialised comparable corpora for travel insurance (Corpas Pastor and Seghiri, 2009), medical treatments (Gutiérrez Florido et al., 2013), among other narrow-domains. WebBootCat[2] (Baroni et al., 2006) is similar to BootCaT, but instead of having to download and install the application, WebBootCat can be used online. Despite being designed for other purposes, Terminus[3] and Corpográfo[4] should also be mentioned as examples of web-based compilation tools.

As we can see, several semi-automatic compilation tools have been proposed so far. Nevertheless, these compilation tools are scarce or proprietary, simplistic with limited features, built to compile one monolingual corpus at a time and do not cover the entire compilation process (i.e. apart from compiling monolingual comparable corpora, they do not allow managing and exploring both parallel and multilingual comparable corpora). Thus, their simplicity, lack of features, performance issues and usability problems result in a pressing need to design new compilation tools tailored to fulfil not only translators' and interpreters' needs, but also professionals' and ordinary people's.

3 iCompileCorpora

Departing from a careful analysis of the weaknesses and strengths of the current compilation solutions, we started by designing and developing a robust and agile web-based application prototype to semi-automatically compile mono- and multilingual comparable corpora, which we named iCompileCorpora. iCompileCorpora can be simply described as a Web graphical interface that will guide the user through the entire corpus compilation process. Designed and implemented from scratch, this application aims to cater to both novice and experts in the field. It will not only provide a simple interface with simplified steps, but also will permit experienced users to set advanced compilation options during the process.

The dimensions that comprise iCompileCorpora can be represented in a layered model comprising a manual, a semi-automatic web-based and a semi-automatic Cross-Language Information Retrieval (CLIR) layer (see Figure 1). This design option will permit not only to increase the flexibility and robustness of the compilation process, but will also hierarchically extend the manual layer features to the

[1] http://bootcat.sslmit.unibo.it
[2] www.sketchengine.co.uk/documentation/wiki/Website/Features#WebBootCat
[3] terminus.iula.upf.edu//cgi-bin/terminus2.0/terminus.pl
[4] http://www.linguateca.pt/corpografo/

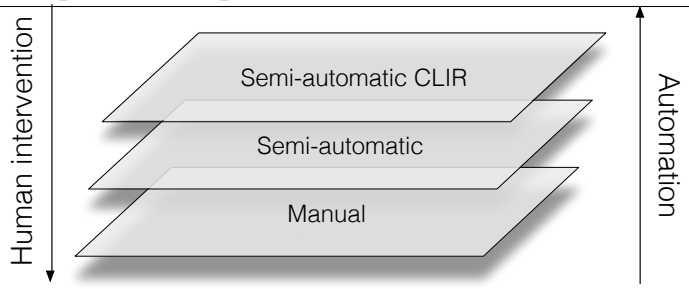


Figure 1: iCompileCorpora layered model.

semi-automatic web-based layer and then to the semi-automatic CLIR layer. In detail, the manual layer represents the option of compiling monolingual and multilingual corpora. It will allow for the manual upload of documents from a local or remote directory onto the platform. The second layer will permit the exploitation of both mono- and multilingual corpora mined from the Internet. Although this layer can be considered similar to the approaches used by BootCaT and WebBootCat, it has been designed to address some of their limitations (e.g. allow the use of more than one boolean operator when creating search query strings), and to improve the User Experience (UX) with this type of software. As nowadays there is an increasing demand for systems that can somehow cross the language boundaries by retrieving information in various languages with just one query, the third layer aims to answer this demand by taking advantage of CLIR techniques to find relevant information written in a language different to the one semi-automatically retrieved by the methodology used in the previous layer.

4 Conclusion

This article presents an ongoing project that aims to increase the flexibility and robustness of the compilation of monolingual and multilingual comparable corpora by creating a new web-based application from scratch. iCompileCorpora intends to fulfil not only translators' and interpreters' needs, but also professionals' and ordinary people's, either by breaking some of the usability problems found in the current compilation tools available on the market or by improving their limitations and performance issues. By the end of this project, we intend to make this compilation tool publicly available, both in a research or in a commercial setting.

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471. Also, the research reported in this work has been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. nº FFI2012-38881, 2012-2015), and the R&D Project for Excelence TERMITUR (ref. nº HUM2754, 2014-2017).

References

Baker, P. (2010). *Sociolinguistic and Corpus Linguistics*. Edinburgh University Press, Edinburgh, UK.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *4th Int. Conf. on Language Resources and Evaluation*, LREC'04, pages 1313–1316.

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *11th Annual Conf. of the European Association for Machine Translation*, EAMT'06, pages 247–252, Oslo, Norway. The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway).

Bowker, L. (2002). *Computer-aided Translation Technology: A Practical Introduction*. Didactics of translation series. University of Ottawa Press.

Bowker, L. and Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Corpas Pastor, G. (2001). Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, 5(1):155–184.

Corpas Pastor, G. and Seghiri, M. (2009). Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). In Beeby, A., Inés, P., and Sánchez-Gijón, P., editors, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014). Technology-assisted Interpreting. *MultiLingual #143*, 25(3):27–32.

EAGLES (1996). Preliminary Recommendations on Corpus Typology. Technical report, EAGLES Document EAG-TCWG-CTYP/P. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

Gries, S. T. (2008). *Corpus-based methods in analyses of SLA data*, pages 406–431. Routledge, NY, USA.

Gutiérrez Florido, R., Corpas Pastor, G., and Seghiri, M. (2013). Using semi-automatic compiled corpora for medical terminology and vocabulary building in the healthcare domain. In *10^{th} Int. Conf. on Terminology and Artificial Intelligence (TIA'13), Workshop on Optimizing Understanding in Multilingual Hospital Encounters*, Paris, France.

Hollmann, W. and Siewierska, A. (2006). Corpora and (the need for) other methods in a study of Lancashire dialect. *Zeitschrift für Anglistik und Amerikanistik*, 1(54):203–216.

Roland, D., Dick, F., and Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57(3):348–379.

Skadiņa, I., Vasiļjevs, A., Skadiņš, R., Gaizauskas, R., Tufiş, D., and Gornostay, T. (2010). Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. In *3^{rd} Workshop on Building and Using Comparable Corpora (BUCC'10)*, pages 6–14, Valletta, Malta.

Wynne, M. (2006). Stylistics: corpus approaches. *Encyclopedia of Language and Linguistics*, 12(2):223–226.

Zanettin, F., Bernardini, S., and Stewart, D. (2003). *Corpora in Translator Education*. Manchester: St. Jerome Publishing.

Costa et al. (2015d)

Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015). **iCorpora: Compiling, Managing and Exploring Multilingual Data**. In *7th Int. Conf. of the Iberian Association of Translation and Interpreting Studies, AIETI*, pages 74-76, Malaga, Spain.

iCorpora: Compiling, Managing and Exploring Multilingual Data

Hernani Costa[a], Gloria Corpas Pastor[a], Miriam Seghiri[a] and Ruslan Mitkov[b]
[a]LEXYTRAD, University of Malaga, Spain
[b]RIILP, University of Wolverhampton, UK
{hercos,gcorpas,seghiri}@uma.es, r.mitkov@wlv.ac.uk

Abstract

In the last decade, there has been a growing interest in bilingual and multilingual corpora. Particularly, in translation their benefits have been demonstrated by several authors (cf. Bowker and Pearson (2002); Bowker (2002); Zanettin et al. (2003); Corpas Pastor and Seghiri (2009)). Their objectivity, reusability, multiplicity and applicability of uses, easy handling and quick access to large volume of data are just an example of their advantages. Thus, it is not surprising that the use of corpora has been considered an essential resource in several research domains such as translation, language learning, stylistics, sociolinguistics, terminology, language teaching, automatic and assisted translation, amongst others. Nevertheless, the lack of sufficient/up-to-date parallel corpora and linguistic resources for narrow domains and poorly-resourced languages is currently one of the major obstacles to further advancement on these areas. One potential solution to the insufficient parallel translation data is the exploitation of non-parallel bilingual and multilingual text resources, also known as comparable corpora (i.e. corpora that include similar types of original texts in one or more language using the same design criteria (cf. EAGLES (1996); Corpas Pastor, 2001:158). Even though comparable corpora can compensate for the shortage of linguistic resources and ultimately improve automated translations quality for under-resourced languages and narrow domains for example, the problem of data collection presupposes a significant technical challenge. The solution proposed in iCorpora project and presented in this article is to exploit the fact that comparable corpora are much more widely available than parallel translation data. This ongoing project aims to increase the flexibility and robustness of the compilation, management and exploration of both comparable and parallel corpora by creating a new web-based application from scratch. iCorpora intends to fulfil not only translators' and interpreters' needs (Costa et al. (2014b;a)), but also professionals' and ordinary people's, either by breaking some of the usability problems found in the current compilation tools available on the market (e.g. BootCaT (Baroni and Bernardini (2004)) and WebBootCat (Baroni et al. (2006)) or by improving their limitations and performance issues. iCorpora will aggregate three applications: iCompileCorpora, iManageCorpora and iExploreCorpora. The first application, iCompileCorpora (Costa et al. (2014c)), can be seen as a layered model comprising a manual, a semi-automatic web-based and a semi-automatic Cross-Language Information Retrieval (CLIR) layer. This design option will permit not only to increase the flexibility and robustness of the compilation process, but will also hierarchically extend the manual layer features to the semi-automatic web-based layer and then to the semi-automatic CLIR layer (i.e. the CLIR layer will automatically translate the queries to other languages (Talvensaari et al. (2007))). iManageCorpora will be specially designed to: manage (i.e. it will allow to edit, copy and paste sentences and documents from and to documents and corpora respectively, as well as to manage corpora into domains and sub-domains); measure the similarity between documents; and to explore the representativeness of the corpora (cf. Corpas Pastor and Seghiri (2009)). Finally, iExploreCorpora intends to offer a set of concordance features, such as search for words in context, automatic extraction of the most frequent words and multi-words, amongst other.

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471. Also, the research reported in this work has been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. nº FFI2012-38881, 2012-2015), and the R&D Project for Excelence TERMITUR (ref. nº HUM2754, 2014-2017).

References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *4th Int. Conf. on Language Resources and Evaluation*, LREC'04, pages 1313–1316.

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *11th Annual Conf. of the European Association for Machine Translation*, EAMT'06, pages 247–252, Oslo, Norway. The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway).

Bowker, L. (2002). *Computer-aided Translation Technology: A Practical Introduction*. Didactics of translation series. University of Ottawa Press.

Bowker, L. and Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Corpas Pastor, G. (2001). Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, 5(1):155–184.

Corpas Pastor, G. and Seghiri, M. (2009). Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). In Beeby, A., Inés, P., and Sánchez-Gijón, P., editors, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014a). A comparative User Evaluation of Terminology Management Tools for Interpreters. In *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, pages 68–76, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014b). Technology-assisted Interpreting. *MultiLingual #143*, 25(3):27–32.

Costa, H., Corpas Pastor, G., and Seghiri, M. (2014c). iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, London, UK.

EAGLES (1996). Preliminary Recommendations on Corpus Typology. Technical report, EAGLES Document EAG-TCWG-CTYP/P. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., and Keskustalo, H. (2007). Creating and Exploiting a Comparable Corpus in Cross-language Information Retrieval. *ACM Transactions on Information Systems*, 25(1).

Zanettin, F., Bernardini, S., and Stewart, D. (2003). *Corpora in Translator Education*. Manchester: St. Jerome Publishing.

Costa et al. (2015a)

Costa, H., Béchara, H., Taslimipoor, S., Gupta, R., Orasan, C., Corpas Pastor, G., and Mitkov, R. (2015). **MiniExperts: An SVM approach for Measuring Semantic Textual Similarity**. In *9th Int. Workshop on Semantic Evaluation, SemEval'15*, pages 96-101, Denver, Colorado. ACL.

MiniExperts: An SVM Approach for Measuring Semantic Textual Similarity

Hernani Costa[*b], **Hanna Béchara**[*a], **Shiva Taslimipoor**[a], **Rohit Gupta**[a], **Constantin Orăsan**[a], **Gloria Corpas Pastor**[b] **and Ruslan Mitkov**[a]
[a]RIILP, University of Wolverhampton, UK
[b]LEXYTRAD, University of Malaga, Spain
{hercos,hanna.bechara,shiva.taslimi,r.gupta,
c.orasan,gcorpas,r.mitkov}@{[a]wlv.ac.uk,[b]uma.es}
[*]These two authors contributed equally to this work.

Abstract

This paper describes the system submitted by the University of Wolverhampton and the University of Malaga for SemEval-2015 Task 2: *Semantic Textual Similarity*. The system uses a Supported Vector Machine approach based on a number of linguistically motivated features. Our system performed satisfactorily for English and obtained a mean 0.7216 Pearson correlation. However, it performed less adequately for Spanish, obtaining only a mean 0.5158.

1 Introduction

Similarity measures play an important role in a wide variety of Natural Language Processing (NLP) applications. Information Retrieval (IR), for example, relies on semantic similarity in order to determine the best result for a related query. Semantic similarity also plays a crucial role in other applications such as Paraphrasing and Translation Memory (TM). However, computing semantic similarity between sentences remains a complex and difficult task. Over the years, SemEval's shared tasks worked to fine-tune and perfect these similarity measures, and explore the nature of meaning in language.

SemEval2015's Task 2 involves computing how similar two sentences are in both English (Subtask 2a) and Spanish (Subtask 2b). In this paper we detail our submission to SemEval Task 2. We use an improved and revised version of the system presented in our SemEval 2014 submission (Gupta et al., 2014). As in Gupta et al., 2014, we employ a Machine Learning (ML) method which exploits available NLP technology, adding features inspired by deep semantics (such as parsing and paraphrasing) with distributional Similarity Measures, Conceptual Similarity Measures, Semantic Similarity Measures and Corpus Pattern Analysis[1] (CPA).

The remainder of the paper is structured as follows. Section 2 describes our approach, i.e. explains how the data was preprocessed and what features were extracted. Section 3 is divided in two section, the first one describes the ML algorithm and how it was tuned for this task (section 3.1) and the second one shows the obtained results along with a descriptive analysis of the runs based on the test and training data provided by the SemEval-2015 Task 2 (section 3.2). Finally, section 4 presents the final remarks and highlights our future plans for improving the system.

2 Approach

This section describes our approach to calculating semantic relatedness. It covers all the required preprocessing steps to extract the features themselves.

2.1 Data Preprocessing

This section presents all the tools, libraries and frameworks used to preprocess not only the test datasets but also the training datasets.

2.1.1 POS-Tagger, Lemmatiser, Stemmer

The software we used for these specific NLP tasks were: the Stanford CoreNLP[2] (Toutanova et al.,

[1] http://pdev.org.uk
[2] http://nlp.stanford.edu/software/corenlp.shtml

2003) toolkit, which provides a lemmatiser, POS-Tagger, NER, parsing, and coreference; the TT4J[3] library, which is a Java wrapper around the popular TreeTagger (Schmid, 1995); and the Porter stemmer algorithm provided by the Snowball[4] library.

2.1.2 Named Entity Recogniser (NER)

The library used to identify named entities in English and Spanish was the Apache OpenNLP library[5]. For English, all the pre-trained NER models made available by the Apache OpenNLP library were used (i.e. we used models to identify dates, locations, money, organisations, percentages, persons and time). We also used all the pre-trained NER models for Spanish (in this case, we used models to identify persons, organisations, locations and miscellanea).

2.1.3 Translation Model

Since one of the features we implemented was available only for English (i.e. the Semantic Similarity Measures), we trained a Statistical Machine Translation (SMT) system to translate our Spanish dataset into English. For this purpose, we used the PB-SMT system Moses (Koehn et al., 2007), 5-gram language models with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002), the GIZA++ implementation of IBM word alignment model 4 (Och and Ney, 2003), with refinement and phrase-extraction heuristics as described in Koehn et al., 2003. We trained this system on the Europarl Corpus (Koehn, 2005) and used Minimum Error Rate Training (MERT) (Och, 2003) for tuning on the development set.

2.1.4 Resources

Given that a number of our features depends on stopwords (see section 2.2), we compiled two lists of stopwords, one for English and another one for Spanish. Both are freely available to download[6].

We also used two lists (English and Spanish) of candidates for Multiword Expressions (MWEs) as a resource for one of the features (see section 2.2.5). These lists were extracted from the Europarl Corpus (Koehn, 2005) using the collocation modules of the NLTK package (Loper and Bird, 2002), and sorted by the degree of likelihood association between their components.

2.2 Extracted Features

This section details the features that our system uses to measure the semantic textual similarity between two sentences. The system uses the same features for both Subtask 2a and Subtask 2b. In addition to the baseline features used in Gupta et al., 2014, we introduced a set of Distributional, Semantic and Conceptual Similarity Measures, as well as a feature reflecting MWEs across sentences.

2.2.1 Baseline Features

The system is built on the baseline system developed for SemEval2014, which consists of 13 features explained in detail in Gupta et al., 2014. The code which implements these features can be found on GitHub[7].

2.2.2 Distributional Similarity Measures

Information Retrieval (IR) (Singhal, 2001) is the task of locating specific information within a collection of documents or other natural language resources according to some request (Salton and Buckley, 1988; Costa et al., 2010; Costa et al., 2011). Among IR methods, we can find a large number of statistical approaches based on the occurrence of words in documents or sentences. Following Harris' distributional hypothesis (Harris, 1970), which assumes that similar words tend to occur in similar contexts, these methods are suitable, for instance, to find similar sentences based on the words they contain or to compute the similarity of words based on their co-occurrence. To that end, we can assume that the amount of information contained in a sentence could be evaluated by summing the amount of information contained in the sentence words. Moreover, the amount of information conveyed by a word can be represented by means of the weight assigned to it (Salton and Buckley, 1988). Bearing this in mind, we used two independent IR measures, the Spearman's Rank Correlation Coefficient (SCC) and the χ^2 to compute the similarity between two sentences

[3] https://code.google.com/p/tt4j
[4] http://snowball.tartarus.org
[5] http://opennlp.apache.org
[6] https://github.com/hpcosta/stopwords
[7] https://github.com/rohitguptacs/wlvsimilarity

written in the same language (cf. Kilgarriff, 2001). Both measures are particularly useful for this task because they are independent of text size (mostly because both measures use a list of the common entities), and they are language-independent. In detail, for every pair of sentence (English and Spanish), we used the lemmas to extract the list of common terms to compute both measures.

2.2.3 Conceptual Similarity Measures

This feature aims to find the conceptual similarity between two sentences written in the same language. In order to calculate the conceptual similarity, we took advantage of the BabelNet[8] (Navigli and Paolo Ponzetto, 2012) multilingual semantic network. As BabelNet organises lexical information in a semantic conceptual way, we created a conceptual sentence for all input pair of sentences (English and Spanish). More precisely, for every pair of sentence we only extracted lemmatised nouns, verbs, adjectives and adverbs. Then, a conceptual term list was built by extracting all the occurrences of the term in the conceptual network (i.e. BabelNet). As a result, we got a "conceptual representation" of both sentences, each of them containing a set of conceptual term lists. Next, for every term in the "conceptual_sentence_1", we counted the number of co-occurrences in the conceptual term lists in the "conceptual_sentence_2". In other words, we intersected the terms in sentence 1 with all the conceptual term lists in sentence 2. After computing all the co-occurrences, we used these values to calculate the Jaccard' (Jaccard, 1901), Lin' (Lin, 1998) and PMI' (Turney, 2001) scores.

2.2.4 Semantic Similarity Measures

This feature takes advantage of the Align, Disambiguate and Walk (ADW)[9] library (Pilehvar et al., 2013), a WordNet-based approach for measuring semantic similarity of arbitrary pairs of lexical items. It is important to mention that this feature is the only one that only works for English, which explains why we have a translation model (see section 2.1.3). In other words, when we are dealing with Spanish text, we use the trained model to translate from Spanish to English.

As the ADW library permits us to measure the semantic similarity between two raw English sentences, either by using disambiguation or not, we used both options to calculate all the comparison methods made available by the library, i.e. WeightedOverlap, Cosine, Jaccard, KLDivergence and JensenShannon divergence.

2.2.5 Multiword Expressions

Multiword Expressions (MWEs) are meaningful lexical units whose distinct idiosyncratic properties call for special treatment within a computational system. Non-compositionality is one of the properties of MWEs. The degree of association between the components of a MWE has been proved to be a promising approach to find out how much they are non-compostional and therefore how probable they are acceptable MWEs (Ramisch et al., 2010). The more non-compositional a MWE is, the more important is not to treat its components separately for NLP purposes, including processing semantic similarities.

For the purpose of our experiments, we focused on two more common types of MWEs in English and Spanish: `verb noun` combinations and `verb particle` constructions. Whenever a `verb+noun` or a `verb+particle` combination occurs in our sentence pair, we search a prepared list MWEs, sorted according to their likelihood measures of association. The degree of association of these combinations served as a feature in our ML system.

3 Predicting Through Machine Learning

In this section, we outline the ML model trained on the extracted features to compute a relatedness score between two sentences. It details the tools and parameters used to build a support vector regressor, which we used to predict a number between 0 and 5, denoting a degree of semantic similarity.

3.1 Model Description

We used a Support Vector Machine (SVM) in order to compute semantic relatedness for both subtasks.

[8] `http://babelnet.org`
[9] `http://lcl.uniroma1.it/adw`

We used LibSVM[10], a library for SVMs developed by Chang and Lin, 2011.

We built a regression model which estimates a continuous score between 0 and 5 for each sentence pair. The values of C and γ have been optimised through a grid-search which uses a 5-fold cross-validation method, and all systems use an RBF kernel.

The system for Subtask 2a (English) is trained on a combination of training and trial data provided by the 2012, 2013 and 2014 SemEval tasks. We used these datasets to form a training set of 9750 sentence pairs combining the different domains covered by the STS task: image description (image), news headlines (headlines), student answers paired with reference answers (answers-students), answers to questions posted in stach exchange forums (answers-forum), English discussion forum data exhibiting committed belief (belief). However, the training set for Subtask 2b (Spanish) was much smaller, at only 804 sentence pairs collected by combining previous datasets from the Newswire and Wikipedia domains.

3.2 Results and Analysis

The task required the submission of 3 different runs for each task. The runs for the Subtask 2a (English) were identical except for some parameter differences for the SVM training. Our system performed adequately, with our primary run achieving a mean Pearson Correlation of 0.7216.

However, the runs for Subtask 2b (Spanish) were trained on different training sets. Run-1 and Run-2 are trained on the 804 Spanish sentence-pairs. The Spanish set's Run-3, however, is trained on the much larger English training set. For this purpose, we needed to translate the Spanish test set into English in order to use the Semantic Similarity language-dependent features (see sections 2.1.3 and 2.2.4). This system did not outperform the basic Spanish model used in Run-1 and Run-2, despite the much larger training set. Our Spanish system did not yield a satisfactory performance, achieving a Pearson Correlation score of only 0.5158. This could be part due to the smaller training set in Spanish, and the imperfect translations into English which consequently influenced the performance of the language-dependent features. The detailed results for both tasks are given in Table 1 and 2.

	Run-1	**Run-2**	**Run-3**
answers-forums	0.6781	0.6454	0.6179
answers-students	0.7304	0.7093	0.6977
belief	0.6294	0.5165	0.3236
headlines	0.6912	0.6084	0.5775
images	0.8109	0.7999	0.7954
mean	***0.7216***	***0.6746***	***0.6353***
rank (out of 74)	***33***	***45***	***55***

Table 1: Task 2a – Pearson Correlation for English.

	Run-1	**Run-2**	**Run-3**
wikipedia	0.5239	0.4671	0.4402
newswire	0.5076	0.5437	0.5524
mean	***0.5158***	***0.5054***	***0.4963***
rank (out of 17)	***9***	***10***	***11***

Table 2: Task 2b – Pearson Correlation for Spanish.

4 Conclusion and Future Work

We have presented an efficient approach to calculate semantic relatedness for both English and Spanish sentence pairs. We used the same feature set for both tasks, even though it meant translating the Spanish sentences into English before extracting one of the features (i.e. the Semantic Similarity). The system did not performed well for Spanish as it ranked 9 out of 17, with a 0.5158 average Person correlation over two test sets (0.1747 correlation points less than the best submitted run). On the other hand, it performed reasonably well for English, where the system's best result ranked 33 among 74 submitted runs with 0.7216 Pearson correlation over five test sets (only 0.0799 correlation points less than the best submitted run).

In the future we plan to extract the conceptual description provided by the BabelNet network in order to match it with the conceptual terms. We have not done that for now because we need to treat these descriptions as sentences, which requires filtering out the noise produced by them.

[10] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Acknowledgements

Hanna Béchara, Hernani Costa and Rohit Gupta are supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471. Also, the research reported in this work has been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. nº FFI2012-38881, 2012-2015); and the R&D Project for Excellence TERMITUR (ref. nº HUM2754, 2014-2017).

References

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2010. The Impact of Distributional Metrics in the Quality of Relational Triples. In *19th European Conf. on Artificial Intelligence, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, ECAI'10, pages 23–29, Lisbon, Portugal.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2011. Using the Web to Validate Lexico-Semantic Relations. In *15th Portuguese Conf. on Artificial Intelligence*, volume 7026 of *EPIA'11*, pages 597–609, Lisbon, Portugal. Springer.

Rohit Gupta, Hanna Bechara, Ismail El Maarouf, and Constantin Orasan. 2014. UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment. In *8th Int. Workshop on Semantic Evaluation (SemEval'14)*, pages 785–789, Dublin, Ireland. ACL and Dublin City University.

Zelig Harris. 1970. Distributional Structure. In *Papers in Structural and Transformational Linguistics*, pages 775–794. D. Reidel Publishing Company, Dordrecht, Holland.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Adam Kilgarriff. 2001. Comparing Corpora. *Int. Journal of Corpus Linguistics*, 6(1):97–133.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Conf. of the North American Chapter of the ACL on Human Language Technology - Volume 1*, NAACL'03, pages 48–54. ACL.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit*.

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. In *15th Int. Conf. on Machine Learning*, ICML'98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP'02, pages 62–69. ACL.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Franz Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *41st Annual Meeting on ACL - Volume 1*, ACL'03, pages 160–167. ACL.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In *51st Annual Meeting of the ACL - Volume 1*, pages 1341–1351, Sofia, Bulgaria. ACL.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. Multiword Expressions in the Wild?: The Mwetoolkit Comes in Handy. In *23rd Int. Conf. on Computational Linguistics: Demonstrations*, COLING'10, pages 57–60. ACL.

Gerard Salton and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

Amit Singhal. 2001. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer*

Society Technical Committee on Data Engineering, 24(4):35–42.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *7^{th} Int. Conf. on Spoken Language Processing*, ICSLP'02, pages 901–904.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *HLT-NAAC 2003*, pages 252–259, Edmonton, Canada. ACL.

Peter D. Turney. 2001. Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. In *12^{th} European Conf. on Machine Learning*, EMCL'01, pages 491–502, London, UK. Springer.

Costa (2015)

Costa, H. (2015). **Assessing Comparable Corpora through Distributional Similarity Measures**. In *EXPERT Scientific and Technological Workshop*, pages 23-32, Malaga, Spain. Tradulex.

Assessing Comparable Corpora through Distributional Similarity Measures

Hernani Costa
LEXYTRAD, University of Malaga, Spain
hercos@uma.es

Abstract

Describing, comparing and evaluating corpora are key issues in corpus-based translation and corpus linguistics for which there is still a notable lack of standards. Bearing this in mind, this paper aims at investigating the use of textual distributional similarity measures in the context of comparable corpora. More precisely, we address the issue of measuring the relatedness between documents by extracting and measuring their common content. For this purpose, we designed and applied a methodology that exploits available natural language processing technology with statistical methods. Our findings showed that using a list of common entities and a simple, yet robust and high performance set of distributional similarity measures was enough to describe and assess the degree of relatedness between the documents in a comparable corpus.

1 Introduction

The use of comparable corpora has been considered an essential resource in several research domains such as Natural Language Processing (NLP), terminology, language teaching, and automatic and assisted translation, amongst others. Nevertheless, an inherent problem to those who deal with comparable corpora in a daily basis is the uncertainty about the data they are dealing with. Indeed, little work has been done on automatically characterising such linguistic resources and attempting a meaningful description of their content is often a perilous task (Corpas Pastor and Seghiri, 2009). Usually, a corpus is given a short description such as "casual speech transcripts" or "tourism specialised comparable corpus". However, such tags will be of little use to those users seeking for a representative and/or high quality domain-specific corpora. Apart from the usual description that comes along with the corpus, like number of documents, tokens, types, source(s), creation date, policies of usage, etc., nothing is said about how similar the documents are. As a result, most of the resources at our disposal are built and shared without deep analysis of their content, and those who use them blindly trust on the people's or research group's name behind their compilation process, without knowing nothing about the relatedness quality of the corpus.

Bearing this in mind, in this work we try to fill this void by taking advantage of several textual distributional similarity measures presented in the literature. First, we selected a specialised corpus about tourism and beauty domain that was manually compiled by researchers in the area of translation and interpreting studies. Then, we designed and applied a methodology that exploits available NLP technology with statistical methods to assess how the documents correlate with each other in the corpus. Our assumption is that the amount of information contained in a document can be evaluated via summing the amount of information contained in the member words. For this purpose, a list of common entities was used as a unit of measurement capable of identifying the amount of information shared between the documents. Our assumption is that this approach will allow us not only to compute the relatedness between documents, but also to describe and characterise the corpus itself.

The remainder of the paper is structured as follows. Section 2 introduces some fundamental concepts related to distributional similarity measures, i.e. explains the theoretical foundations, related work and the distributional similarity exploited in this experiment. Then, Section 3 presents the corpus used in this work. After applying the methodology described in Section 4, Section 5 presents and discusses the

obtained results in detail. Finally, Section 6 presents the final remarks and highlights our future plans for this work.

2 Distributional Similarity Measures

Information Retrieval (IR) (Singhal, 2001) is the task of locating specific information within a collection of documents or other natural language resources according to some request. In this field, we can find a large number of statistical methods based on words and their (co-)occurrence. Essentially, it involves finding the most frequently used words and treating the rate of usage of each word in a given text as a quantitative attribute. Then, these words serve as features for a given statistical method. Following Harris' distributional hypothesis (Harris, 1970), which assumes that similar words tend to occur in similar contexts, these statistical methods are suitable, for instance to find similar sentences based on the words they contain (Costa et al., 2015a) and automatically extract or validate semantic entities from corpora (Costa et al., 2010; Costa, 2010; Costa et al., 2011). To this end, it is assumed that the amount of information contained in a document could be evaluated by summing the amount of information contained in the document words. And, the amount of information conveyed by a word can be represented by means of the weight assigned to it (Salton and Buckley, 1988). Accordingly, we took advantage of two IR measures commonly used in the literature, the Spearman's Rank Correlation Coefficient (SCC) and the Chi-Square (χ^2) to compute the similarity between two documents written in the same language (see section 2.1 and 2.2). Both measures are particularly useful for this task because they are independent of text size (mostly because both use a list of the common entities), and they are language-independent.

The Spearman's Rank Correlation Coefficient (SCC) distributional measure has been shown effective on determining similarity between sentences, documents and even on corpora of varying sizes (Kilgarriff, 2001; Costa et al., 2015a). It is particularly useful, for instance to measure the textual similarity between two documents because it is easy to compute and is independent of text size as it can directly compare ranked lists for large and small texts.

The χ^2 similarity measure has also shown its robustness and high performance. By way of example, χ^2 have been used to analyse the conversation component of the British National Corpus (Rayson et al., 1997), to compare corpora (Kilgarriff, 2001), and to identify topic related clusters in imperfect transcribed documents (Ibrahimov et al., 2002). It is a simple statistic measure that permits to assess if relationships between two variables in a sample are due to chance or the relationship is systematic.

For all these reasons, distributional similarity measures in general and SCC and χ^2 in particular have a wide range of applicabilities (cf. Kilgarriff (2001) and Costa et al. (2015a)). Indeed, this work aims at proving that these simple, yet robust and high-performance measures allow to describe the relatedness between documents in specialised corpora.

2.1 Spearman's Rank Correlation Coefficient (SCC)

In this work, the SCC is adopted and calculated as in Kilgarriff (2001). Firstly, a list of the common entities[1] L between two documents d_l and d_m is compiled, where $L_{d_l,d_m} \subseteq (d_l \cap d_m)$. It is possible to use the top n most common entities or all common entities between two documents, where n corresponds to the total number of common entities considered $|L|$, i.e. $\{n|n \in \mathbb{N}^0, n \leq |L|\}$ – in this work we use all the common words for each document pair, i.e. $n = |L|$. Then, for each document the list of common entities (e.g. L_{d_l} and L_{d_m}) is ranked by frequency in an ascending order ($R_{L_{d_l}}$ and $R_{L_{d_m}}$), where the entity with lowest frequency receives the numerical raking position 1 and the entity with highest frequency receives the numerical raking position n. In the case of ties in rank, where more than one entity in a document occurs with the same frequency, the average of the ranks is assigned to the tying entities. For instance, if the entities e_a, e_b and e_c had the same frequency and ranked in the 6^{th}, 7^{th} and 8^{th} position, all three entities would be assigned the same rank of $\frac{6+7+8}{3} = 7$. Finally, for each common entity $\{e_1, ..., e_n\} \in L$, the difference in the rank orders for the entity in each document is computed,

[1] In this work, the term 'entity' refers to "single words", which can be a token, a lemma or a stemm.

and then normalised as a sum of the square of these differences $\left(\sum_{i=1}^{n} s_i^2\right)$. The final SCC equation is presented in expression 1, where $\{SCC|SCC \in \mathbb{R}, -1 \geq SCC \leq 1\}$.

By a way of example let e_x be a common entity (i.e. $\{e_x\} \in L$) and $R_{L_{d_l}} = \{1\#e_{n_{d_l}}, 2\#e_{n-1_{d_l}}, ..., n\#e_{1_{d_l}}\}$ and $R_{L_{d_m}} = \{1\#e_{n_{d_m}}, 2\#e_{n-1_{d_m}}, ..., n\#e_{1_{d_m}}\}$ the resulting ranked list of common words for d_l and d_m, respectively. Supposing that e_x is the $3\#e_{n-2_{d_l}}$ and $1\#e_{n_{d_m}}$, i.e. e_x is in the 3^{rd} position in $R_{L_{d_l}}$ and in the 1^{st} position in $R_{L_{d_m}}$, s would be computed as $s_{e_x}^2 = (3-1)^2$ and the result would be 4. Then, this process is repeated for the remain $n-1$ entities and the resulted SCC score will be seen as the similarity value between d_l and d_m.

$$SCC(d_i, d_j) = 1 - \frac{6 * \sum_{i=1}^{n} s_i^2}{n^3 - n} \tag{1}$$

2.2 Chi-Square (χ^2)

The Chi-square (χ^2) measure also uses a list of common words (L). Similarly to SCC, it is also possible to use the top n most common entities or all common entities between two documents, and again in this work we use all the common words for each document pair, i.e. $n = |L|$. The number of occurrences of a common words in L that would be expected in each document is calculated from the frequency lists. If the size of the document d_l and d_m are N_l and N_m and the entity e_i has the following observed frequencies $O(e_i, d_l)$ and $O(e_i, d_m)$, then the expected values are $e_{i_{d_l}} = \frac{N_l * (O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$ and $e_{i_{d_m}} = \frac{N_m * (O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$. Equation 2 presents the χ^2 formula, where O is the observed frequency and E the expected frequency. The resulted χ^2 score should be interpreted as the interdocument distance between two documents. It is also important to mention that $\{\chi^2|\chi^2 \in \mathbb{R}, 1 \geq \chi^2 < \infty\}$, which means that as more unrelated the common words in L are, the lower the χ^2 score will be.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{2}$$

Suppose that we have two common entities e_i and e_j between two documents d_l and d_m (i.e. $L = \{e_i, e_j\}$). Table 1 shows a contingency table example. This table contains: i) the observed frequencies (O); ii) the totals in the margins; iii) and the expected frequencies (E), which are obtained by applying the following formula: $\frac{column\ total}{N} * row\_total$, e.g. $E(e_i, d_l) = \frac{14}{26} * 15 = 8.08$. After writing down the expected frequencies in the table, we are ready to calculate the χ^2 score (see Equation 3).

	d_l	d_m	*Total*
e_i	O=11 E=8.08	O=4 E=6.92	15
e_j	O=3 E=5.92	O=8 E=5.08	11
Total	14	12	26

Table 1: Example of a contigency table.

$$\chi^2 = \frac{(11-8.08)^2}{8.08} + \frac{(3-5.92)^2}{5.92} + \frac{(4-6.92)^2}{6.92} + \frac{(8-5.08)^2}{5.08} = 5.41 \tag{3}$$

3 The INTELITERM Corpus

The INTELITERM[2] corpus is a comparable corpus composed of documents collected from the Internet. Designed to be a specialised comparable corpus, this corpus was manually compiled by researchers

[2] http://www.lexytrad.es/proyectos.html

with the purpose of building a representative corpus for the Tourism and Beauty domain. It contains documents in four different languages (English, Spanish, German and Italian). Some of the texts are translations of each other, yet the majority is composed of original texts. The INTELITERM comparable corpus is composed of several subcorpora, divided by the language and further for each language there are translated and original texts (which will be hereafter referred as language_totd and language_to, respectively). In this work, we used half of the corpus, i.e. all the original and translated documents in English and Spanish (en_to, en_totd, es_to and es_totd, respectively). All the information about these subcorpora is presented in Table 2. In detail, this table shows: the number of documents (nDocs); the number of types (types); the number of tokens (tokens); and the ratio of types per tokens ($\frac{types}{tokens}$) per subcorpus. These values were obtained using the corpus analysis toolkit for concordancing and text analysis software Antconc 3.4.3 (Anthony, 2014).

	nDocs	types	tokens	$\frac{types}{tokens}$	description
en_to	151	11,6k	508,9k	0.023	original
en_totd	61	6,9k	88,5k	0.078	translated
es_to	225	12,6k	253,4k	0.049	original
es_totd	27	3,4k	19,7k	0.174	translated

Table 2: Statistical information about the various subcorpus.

4 Methodology

This section not only describes the methodology used to calculate the similarity between documents using Distributional Similarity Measures (DSMs), but also presents all the tools, libraries and frameworks employed by our system to perform this experiment.

1) **Data Preprocessing**: firstly all the documents within the corpus were processed with the OpenNLP[3] Sentence Detector and Tokeniser. Then, the annotation process was done with the TT4J[4] library, which is a Java wrapper around the popular TreeTagger (Schmid, 1995) – a tool specifically designed to annotate text with part-of-speech and lemma information. Regarding the stemming, we used the Porter stemmer algorithm provided by the Snowball[5] library. A method to remove punctuation and special characters within the words was also implemented. Finally, in order to get rid of the noise, a stopword list[6] was compiled to filter out the most frequent words in the corpus. Once a document is computed and the sentences are tokenised, lemmatised and stemmed, our system creates a new output file with all this new information, i.e. the new document contains: the original, the tokenised, the lemmatised and the stemmed text. Using the stopword list mentioned above a Boolean vector describing if the entity is a stopword or not is also added. This way, the system will be able to use only the tokens, lemmas and stems that are not stopwords.

2) **Identifying the list of common entities between documents**: in order to identify a list of common entities (L), a co-occurrence matrix was built for each pair of documents. Only those that have at least one occurrence in both documents are considered. As required by the DSMs (see section 2), their frequency in both documents is also stored within this matrix ($L_{d_l,d_m} = \{e_i, (f(e_i,d_l), f(e_i,d_m)); e_j, (f(e_j,d_l), f(e_j,d_m)); ...; e_n, (f(e_n,d_l), f(e_n,d_m))\}$). With the purpose of analysing and comparing the performance of different DSMs, three different lists were created to be used as input features: the first one using common tokens, another using common lemmas and the third one using common stems.

3) **Computing the similarity between documents**: the similarity between documents was calculated by applying three different DSMs ($DSMs = \{DSM_{NCE}, DSM_{SCC}, DSM_{\chi^2}\}$, where $_{NCE}$, $_{SCC}$ and

[3] https://opennlp.apache.org
[4] http://reckart.github.io/tt4j/
[5] http://snowball.tartarus.org
[6] Freely available to download through the following URL https://github.com/hpcosta/stopwords.

χ^2 means Number of Common Entities, Spearman's Rank Correlation Coefficient and Chi-Square, respectively), each one calculated using three different input features (list of common tokens, lemmas and stems).

4) **Computing the document final score**: the document final score $DSM(d_l)$ is the mean of the similarity scores of the document with all the documents in the collection of documents, i.e. $DSM(d_l) = \frac{\sum_{i=1}^{n-1} DSM_i(d_l,d_i)}{n-1}$, where n corresponds to the total number of documents in the collection and $DSM_i(d_l, d_i)$ the resulted similarity score between the document d_l with all the documents in the collection.

5 Results and Analysis

In order to describe the corpus in hand, we applied three different Distributional Similarity Measures (DSMs): the Number of Common Entities (NCE), the Spearman's Rank Correlation Coefficient (SCC) and the Chi-Square (χ^2). As a input feature to the DSMs, three different types of entities (tokens, lemmas and stems) were used. Table 3 shows the Number of Common Tokens (NCT) between document on average (av), the SCC and the χ^2 scores along with the associated standard deviations (σ) per measure and subcorpus. Figure 1 presents the resulted average scores per document in a box plot format for all the combinations DSM vs. feature. Each box plot displays the full range of variation (from min to max), the likely range of variation (the interquartile range), the median, and the high maximums and low minimums (also know as outliers). It is important to mention that for this experiment we did not use a sample, but instead the entire corpus in its original size and form, which means that all obtained results and made observations came from the entire population, in this case the various INTELITERM English (en_to and en_totd) and Spanish (es_to and es_totd) subcorpora.

		NCT	SCC	χ^2
en_to	av	163.70	0.42	279.39
	σ	83.87	0.05	177.45
en_totd	av	67.54	0.39	90.38
	σ	35.35	0.05	53.25
es_to	av	31.97	0.41	40.92
	σ	23.48	0.07	38.21
es_totd	av	17.93	0.63	13.40
	σ	8.46	0.14	18.95

Table 3: Average and standard deviation of common tokens scores between document per subcorpus.

The first observation we can make from Figure 1 is that the distributions between the features are quite similar (see for instance Figures 1a, 1d and 1g). This means that it is possible to achieve acceptable results only using raw words (i.e. tokens). Stems and lemmas require more processing power and time to be used as features – especially lemmas due to the Part-of-Speech (POS) tagger dependency and time consuming process implied. In general, we can say that the scores for each subcorpus is symmetric (roughly the same on each side when cut down the middle), which means that the data is normally distributed. There are some exceptions such as the SCC and χ^2 average scores for the es_totd and for the en_to, respectively, which we will discuss later in this section. Another interesting observation is related with the high NCE (see Table 3 and Figures 1a, 1d and 1g) in original documents (en_to and es_to) when compared with documents translated from other languages (en_totd and es_totd, respectively). For example, the subcorpus en_to (which contains original documents) has 163.70 common tokens per document on average (av) with a standard deviation (σ) of 83.87 and the subcorpus en_totd (which contains translated documents) only has 67.54 common tokens per document on average with a σ=35.35 (Table 3). The same observation can be made between the es_to and the es_totd subcorpus (see Figure 1a and Table 3). This fact could happen because these documents are collections of translated documents collected from the Internet, and thus translated from different translator, which implies that different translators use different vocabulary and consequently lower the NCE between the documents will be.

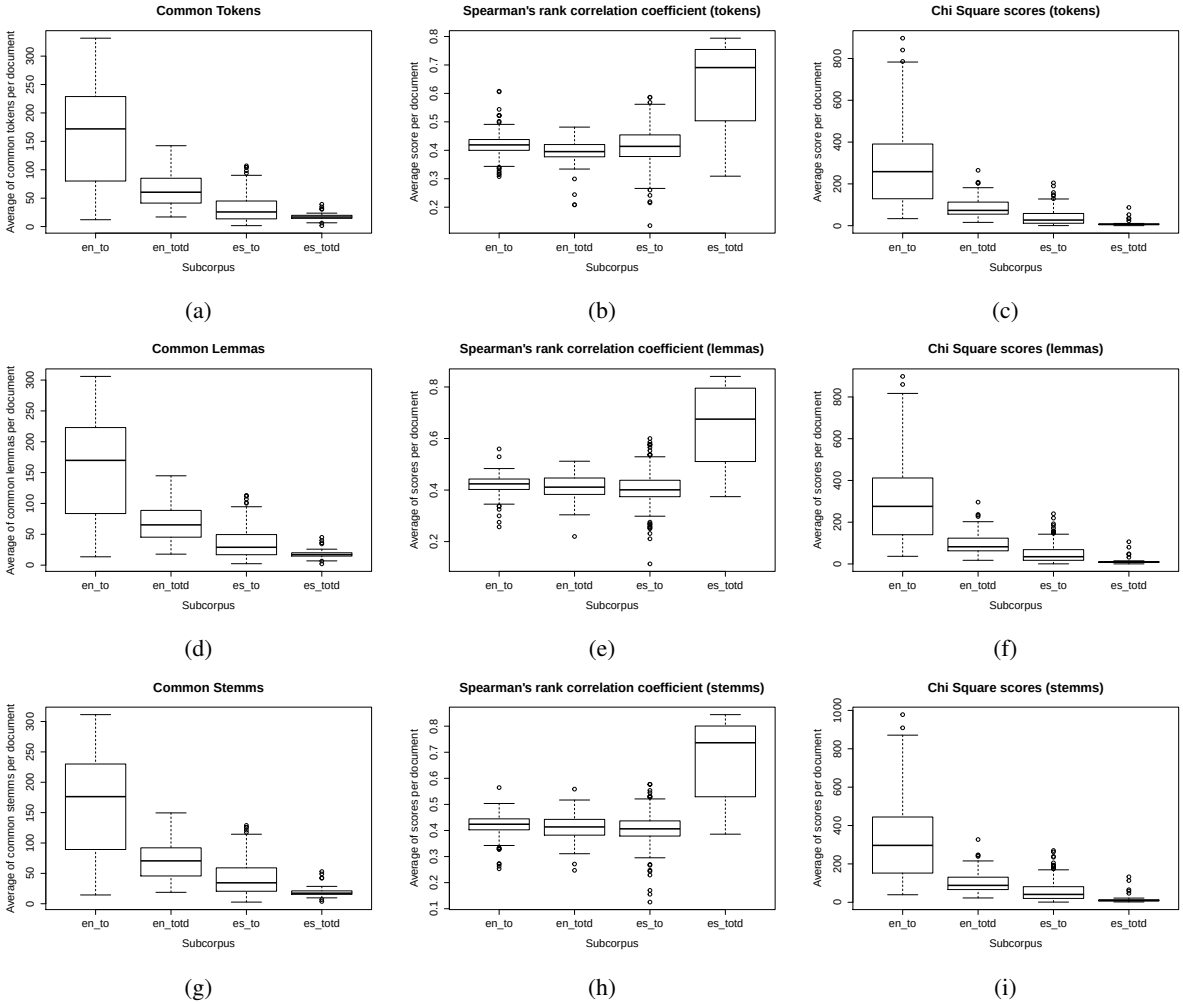


Figure 1: INTELITERM Subcorpus: average scores per document.

Although the Number of Common Tokens (NCT) per document on average is higher for the en_to subcorpus, the interquartile range (IQR) is larger than for the other subcorpora (see Table 3 and Figure 1a), which means that the middle 50% of the data is more distributed and thus the average of NCT per document is more variable. Moreover, longest whiskers (the lines extending vertically from the box) in Figure 1a also indicates variability outside the upper and lower quartiles. Therefore, we can say that en_to has a wide type of documents and consequently some of them are only roughly correlated to the rest of the subcorpus. Nevertheless, the data is skewed right, which means that the majority is strongly similar, i.e. the documents have a high degree of relatedness between each other. This idea can be sustained by the positive average SCC scores presented in Figure 1b and the set of outliers found above the upper whisker. Moreover, the average of 0.42 SCC score and σ=0.05 also implies a strong correlation between the documents in the en_to subcorpus. Likewise, the longest whisker outside the upper quartile and the skewed left χ^2 scores also indicate relatedness between the documents.

Regarding the en_totd subcorpus, the NCT, the SCC and the χ^2 scores (Figures 1a, 1b and 1c) and the average of 90.38 common tokens per document and σ=53.25 (Table 3) suggest that the data is either normally distributed (Figure 1b) or skewed left (Figures 1a and 1c). Considering this results, we can conclude that the documents are highly related.

From all the subcorpora, the es_to subcorpus is the biggest one with 225 documents, 12606 types, 253412 tokens (Table 2). Nevertheless, Table 3 and Figure 1a reveal a lower NCT compared with en_to and the en_totd subcorpora. A theoretical explanation for this phenomenon is that Spanish has richer morphology compared to English. Therefore, due to bigger number of inflection forms per lemma, there

is a larger number of tokens and consequently less common tokens per document in Spanish. When analysing Figures 1a and 1c, the box plots for the es_to subcorpus look similar to the en_totd when shifted up. Except for the longest whisker observed in Figure 1b, the SCC scores also show similar distributions, averages and standard deviations (see Table 3).

As we can see in Figures 1a, 1b and 1c, the average scores per document for es_totd are slightly different from the other box plots. Apart from the low NCT per document, the χ^2 standard deviation higher than its average (18.95 and 13.40, respectively), the SCC variability inside and outside the IQR indicates some inconsistency in the data. This instability can be explained by the subcorpus size, i.e. the small number of documents (27) and by the the low number of types and tokens (3433 and 19736, respectively) and its 0.174 $\frac{types}{tokens}$ ratio. As mentioned by Baker (2006:52), the $\frac{types}{tokens}$ ratio tends to be useful when looking at relatively small documents, and in this specific case this subcorpus only has on average 731 tokens ($\frac{19736}{27} \approx 731$) and 127 types per document ($\frac{3433}{27} \approx 127$), which makes it an excellent test case. When compared with the low ratios from the other subcorpora (see Table 2), – even for this specilised subcorpus – this one can be considered high. If by on one hand, a low ratio can indicate a great number of repetitions (the same word occurring again and again) likely indicating a relatively narrow range of subjects. On the other hand, a high ratio suggests that a more diverse form of language is employed, which can also explain the low NCT and χ^2 scores for this subcorpus in hand. Despite the high SCC, the data is asymmetric and variable (large IQR). This happens because most of the common entities have a low frequency in the documents and consequently they will rank close together in the ranking lists, which results in high SCC scores mostly because of the resulted high value in the numerator (see Equation 1).

To sum up, we can state from the statistical and theoretical evidences that the en_to, the en_totd and the es_to subcorpora look like they assemble highly correlated documents. We can not say the same for the es_totd subcorpus. Due to the small number of documents and scarceness of evidences we can only not reject the idea that this subcorpus is composed of similar documents.

6 Conclusions and Future Work

In this paper we presented and studied various Distributional Similarity Measures (DSMs) for the purpose of describing specialised comparable corpora. As input for these DSMs, we used three different features (lists of common tokens, lemmas and stems). In the end, we conclude that for the data in hand these features had similar performance for all the tested DSMs. In fact, our findings show that instead of using common lemmas or stems, which require external libraries, processing power and time, a simple list of common tokens was enough to describe our data. Moreover, we proved that the corpus used in this experiment is composed of highly correlated documents. The high number of entities shared by its documents, the positive average scores obtained with the SCC measure and their χ^2 scores sustain our claim.

In the immediate future, we intend not only to perform more experiments with these DSMs by adding noisy documents (i.e. out of topic documents) to the corpus and analyse the DSMs performance, but also merge the translated documents from other languages with original ones and prove that translated documents decrease the general relatedness score. Moreover, it is our intention to do the same experiment with other languages, like Italian and German. Apart from that, we also want to test other DSMs, such as Jaccard, Lin and PMI and compare their performance.

Furthermore, these DSMs can be seen as a suitable tool to rank documents by their similarities, which we believe that will be a handy feature to those who manually or semi-automatically compile corpora mined from the Internet. It will allow them to filter out documents with a low level of relatedness when compared with the rest of the documents in the corpus. Indeed, it is our intention to integrate this methodology in the iCorpora application, an ongoing project that aims to design and develop a robust and agile web-based application capable of semi-automatically compile multilingual comparable and parallel corpora (Costa et al., 2014; Costa et al., 2015c; Costa et al., 2015b).

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471. Also, the research reported in this work has been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. nº FFI2012-38881, 2012-2015); and the R&D Project for Excellence TERMITUR (ref. nº HUM2754, 2014-2017). I would like to thank Prof. Gloria Corpas Pastor, Prof. Ruslan Mitkov and Dr. Miriam Seghiri for their valuable comments and suggestions to improve the paper.

References

Laurence Anthony. 2014. AntConc (Version 3.4.3) Machintosh OS X. Waseda University. Tokyo, Japan. Available from `http://www.laurenceanthony.net`.

Paul Baker. 2006. *Using Corpora in Discourse Analysis*. Bloomsbury Discourse. Bloomsbury Academic.

Gloria Corpas Pastor and Míriam Seghiri. 2009. Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). In A. Beeby, P.R. Inés, and P. Sánchez-Gijón, editors, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2010. The Impact of Distributional Metrics in the Quality of Relational Triples. In *19^{th} European Conf. on Artificial Intelligence, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, ECAI'10, pages 23–29, Lisbon, Portugal, August.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2011. Using the Web to Validate Lexico-Semantic Relations. In *15^{th} Portuguese Conf. on Artificial Intelligence*, volume 7026 of *EPIA'11*, pages 597–609, Lisbon, Portugal, October. Springer.

Hernani Costa, Gloria Corpas Pastor, and Miriam Seghiri. 2014. iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, London, UK, November.

Hernani Costa, Hanna Béchara, Shiva Taslimipoor, Rohit Gupta, Constantin Orasan, Gloria Corpas Pastor, and Ruslan Mitkov. 2015a. MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. In *9^{th} Int. Workshop on Semantic Evaluation*, SemEval'15, pages 96–101, Denver, Colorado, June. ACL.

Hernani Costa, Gloria Corpas Pastor, Ruslan Mitkov, and Miriam Seghiri. 2015b. $(In\ press)$ Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora. In *7^{th} Int. Conf. of the Iberian Association of Translation and Interpreting Studies*, AIETI, Malaga, Spain.

Hernani Costa, Gloria Corpas Pastor, Miriam Seghiri, and Ruslan Mitkov. 2015c. iCorpora: Compiling, Managing and Exploring Multilingual Data. In *7^{th} Int. Conf. of the Iberian Association of Translation and Interpreting Studies*, AIETI, pages 74–76, Malaga, Spain, January.

Hernani Costa. 2010. Automatic Extraction and Validation of Lexical Ontologies from text. Master's thesis, University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering, Coimbra, Portugal, September.

Zelig Harris. 1970. Distributional Structure. In *Papers in Structural and Transformational Linguistics*, pages 775–794. D. Reidel Publishing Company, Dordrecht, Holland.

Oktay Ibrahimov, Ishwar Sethi, and Nevenka Dimitrova. 2002. The Performance Analysis of a Chi-square Similarity Measure for Topic Related Clustering of Noisy Transcripts. In *16^{th} Int. Conf. on Pattern Recognition*, volume 4, pages 285–288. IEEE Computer Society.

Adam Kilgarriff. 2001. Comparing Corpora. *Int. Journal of Corpus Linguistics*, 6(1):97–133.

Paul Rayson, Geoffrey Leech, and Mary Hodges. 1997. Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. *Int. Journal of Corpus Linguistics*, 2(1):133–152.

Gerard Salton and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

Amit Singhal. 2001. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42.

Zampieri et al. (2015)

Zampieri, M., Gebrekidan Gebre, B., Costa, H., and van Genabith, J. (2015). **Comparing Approaches to the Identification of Similar Languages**. In *Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial'15), 2nd Discriminating between Similar Languages Shared Task (DSL'15)*, pages 7, Hissar, Bulgaria.

Comparing Approaches to the Identification of Similar Languages

Marcos Zampieri[1,2], Binyam Gebrekidan Gebre[3], Hernani Costa[4] and Josef van Genabith[1,2]
Saarland University, Germany[1]
German Research Center for Artificial Intelligence (DFKI), Germany[2]
Max Planck Computing and Data Facility, Germany[3]
University of Malaga, Spain[4]

Abstract

This paper describes the submission made by the MMS team to the Discriminating between Similar Languages (DSL) shared task 2015. We participated in the closed submission track using only the dataset provided by the shared task organisers which contained short texts from 13 similar languages and language varieties. We submitted three runs using different systems and compare their performance. As a result, our best system achieved 95.24% accuracy for test set A (containing original texts) and 92.78% accuracy for test set B (containing texts without named entities).

1 Introduction

Automatic language identification is an important task in Natural Language Processing (NLP), which consists of applying computational methods to identify the language a document is written in. Language identification is often modelled as a classification task and it is often the first processing stage of many NLP applications and pipelines. Although language identification is largely considered to be a solved task, recent studies have shown that language identification systems often fail to achieve satisfactory performance across different datasets and domains (Lui and Baldwin, 2011), particularly with: datasets containing short pieces of texts such as *tweets* (Zubiaga et al., 2014); code-switching data (Solorio et al., 2014); or when discriminating between very similar languages (Zampieri et al., 2014).

Given these challenges, the Discriminating between Similar Languages (DSL) shared task provides an excellent opportunity for researchers interested in evaluating and comparing their systems' performance on discriminating between similar languages and language varieties using short text excerpts extracted from journalistic texts. For this purpose, the MMS[1] team developed three systems for the closed submission track of the DSL shared task 2015. The systems are explained in more detail in Section 4.

The remainder of the paper is structured as follows. First, Section 2 presents the most relevant approaches in the field. The DSL shared task 2015 is described in detail in Section 3. Then, our approach and the results obtained are presented in Sections 4 and 5 Finally, Section 6 presents the final remarks and highlights our future plans for improving the systems.

2 Related Work

There have been a number of papers published about the identification or discrimination of similar languages in recent years. Most of them use supervised classification algorithms and words and characters as features to solve the task. Unlike general-purpose language identification, most of the systems trained to discriminate between similar languages perform best using high order character n-grams and word n-gram representations.

Different groups or pairs of similar languages and language varieties have been studied using data from different sources such as standard contemporary newspapers and social media. Recent studies include: Indian languages (Murthy and Kumar, 2006), Malay and Indonesian (Ranaivo-Malançon, 2006), Mainland, Singapore and Taiwanese Chinese (Huang and Lee, 2008), Brazilian and European Portuguese (Zampieri and Gebre, 2012), South Slavic languages (Tiedemann

[1]MMS is an acronym for our affiliations/locations (Malaga, Munich and Saarland). In the shared task report (Zampieri et al., 2015) the team is displayed as MMS*. The * indicates that a shared task organiser is a team member.

and Ljubešić, 2012; Ljubešić and Kranjčić, 2015) English varieties (Lui and Cook, 2013), Spanish varieties (Zampieri et al., 2013; Maier and Gómez-Rodrıguez, 2014), and Persian and Dari (Malmasi and Dras, 2015).

Over the last few years there has been a significant increase of interest in the computational processing of Arabic. This is evidenced by a number of research papers on different NLP tasks and applications including the identification/discrimination of Arabic dialects (Elfardy and Diab, 2014; Zaidan and Callison-Burch, 2014; Tillmann et al., 2014; Sadat et al., 2014; Salloum et al., 2014; Malmasi et al., 2015). From a purely engineering perspective, discriminating between dialects poses the same challenges as the discrimination between similar languages and language varieties.

3 The DSL Task

The shared task organisers provided all participants with an updated version of the DSL corpus collection v.2.0 (DSLCC) (Tan et al., 2014). This corpus is composed of 14 classes, 13 languages[2] and one class containing documents written in previously 'unseen' languages to emulate a real-world language identification scenario. Table 1 presents the languages included in the DSLCC v.2.0 corpus grouped by similarity.

Language/ Variety	Code
Bosnian	*bs*
Croatian	*hr*
Serbian	*sr*
Indonesian	*id*
Malay	*my*
Czech	*cz*
Slovak	*sk*
Brazilian Portuguese	*pt-BR*
European Portuguese	*pt-PT*
Argentine Spanish	*es-AR*
Castilian Spanish	*es-ES*
Macedonian	*bg*
Bulgarian	*mk*
Unknown	*xx*

Table 1: DSL corpus by language and variety.

In detail, the corpus collection contains 308,000 short text excerpts sampled from journalistic texts (22,000 per class) varying between 20 and 100 tokens per excerpt.

It is important to mention that these 22,000 texts per class are divided into 3 partitions, i.e. 18,000, 2,000 and 2,000 instances for training, development and testing, respectively. The test set is further subdivided into two test sets (A and B), each one containing 1,000 instances. While the test set A contains original texts, the organisers replaced named entities for place holders in the set B in order to decrease thematic bias in the classification process. Below we present an example of a Portuguese instance containing place holders *#NE#* instead of the named entities.

(1) Compara #NE# este sistema às indulgências vendidas pelo #NE# na #NE# #NE# quando os fiéis compravam a redenção das suas almas dando dinheiro aos padres.

Regarding the choice of only participating in the closed submission track, we first analysed the results of the 2014 edition where we realised that only two teams decided to participate in both open and closed submission tracks, namely UMich (King et al., 2014) and UniMelb-NLP (Lui et al., 2014). Both of them had better performance in the closed submission track and reported that more training data does not necessarily lead to higher performance and that the features learned by the classifiers are, to a certain extent, dataset specific. Therefore, we decided to use only the dataset provided by the organisers and only participate in the closed submission track.

4 Approach

Given that each team was allowed to submit a maximum of three runs to each track (closed and open), we decided to take this opportunity to test and compare different approaches. To do that, we developed three systems based on team MMS-member's previous work in language identification and related tasks. The first two systems were previously used for the Native Language Identification (NLI) (Gebre et al., 2013) and the third one has been applied to language variety identification. The following is a list of the three systems and the their corresponding submission runs:

- **Run 1** - Logistic Regression with TF-IDF Weighting

[2]For the sake of simplicity, we refer to both languages and language varieties as languages.

- **Run 2** - SVM with TF-IDF Weighting
- **Run 3** - Likelihood Estimation

It is important to mention that in each run we used different groups of features, all of them based on n-grams. In detail, for *Run 1* and *Run 2* we used n-grams ranging from bi- to seven-grams and 5-grams for *Run 3*.

4.1 TF-IDF Weighting

Term Frequency - Inverse Document Frequency (TF-IDF)[3] weighting measure was used in the systems developed for *Run 1* and *Run 2*.

Term Frequency refers to the number of times a particular term appears in a text.[4] It seems intuitive to think that a term that occurs more frequently tends to be a better identifier for the text than a term that occurs less frequently, however, this intuition does not take into account the relationship between the frequency of a term and its importance to the text. For this reason, we computed a logarithmic relationship (sublinear TF scaling) (Manning et al., 2008):

$$wf_{t,d} = \begin{cases} 1 + \log(tf_{t,d}) & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $wf_{t,e}$ refers to weight and $tf_{t,e}$ refers to the frequency of term t in document d.

The $wf_{t,d}$ weight represents the importance of a term in a document based on its frequency. However, not all terms that occur frequently in a text are equally important for our purpose. As an example, lets suppose we need to train a classifier to distinguish between British and American English varieties. Words like *the, of, and* will be very frequent, but they are not discriminative, mostly because they are frequent in both varieties. On the other hand, words like *London* or *rubbish* might not be as frequent as *the, of, and*, yet, they are better discriminative words for British English. Therefore, the actual importance of a term for this task depends on how infrequent the term is in other texts. This can be modelled using Inverse Document Frequency (IDF). IDF is based on the assumption that a term which occurs in many texts is not a good discriminator, and should be given less weight than one which occurs in fewer texts. To summarize, IDF is the *log* of the inverse probability of a term being found in any document (Salton and McGill, 1986):

$$idf(t_i) = \log \frac{N}{n_i} \quad (2)$$

where N is the number of documents in the corpus, and term t_i occurs in n_i of them.

TF gives more weight to a frequent term in a document whereas IDF decreases this weight if the term occurs in many documents. On their own, these measures are not very powerful as when combined together to form the well-known TF-IDF measure. The TF-IDF formula combines the weights of TF and IDF by multiplying them. Returning to our example, *the* is a frequent English word so its TF value will be high, however, it is a frequent word in all English texts, in turn making its IDF value low.

Equation 3 shows the final weight that each term in a document gets before normalisation.

$$w_{i,d} = (1 + \log(tf_{t,d})) \times \log \frac{N}{n_i} \quad (3)$$

The texts included in the shared task dataset have different lengths ranging between 20 and 100 tokens each. To cope with this variation we normalised each document feature vector to unit length so that document length does not severely impact term weights. The resulting document feature vectors are fed into two different classifiers, Logistic Regression and SVM.

4.2 Classifiers

Systems developed for *Run 1* and *Run 2* were previously used in the Native Language Identification (NLI) (Gebre et al., 2013) shared task 2013 (Tetreault et al., 2013) by the Cologne-Nijmegen team with good results. They both rely on the TF-IDF weighting scheme combined with two different classifiers.

For *Run 1*, we opt for Logistic Regression using the LIBLINEAR open source library (Fan et al., 2008) from scikit-learn (Pedregosa et al., 2011) and fix the regularisation parameter to 100.0. This regression algorithm has been used in different classification problems including for example temporal text classification (Niculae et al., 2014).

[3] The TF-IDF description presented in this section is based on our previous work (Gebre et al., 2013)

[4] In our experiments, terms are n-grams of characters, words, part-of-speech tags or any combination of them.

For *Run 2*, we used a Support Vector Machine classifier (Joachims, 1998). This approach delivered a slightly better performance than Logistic Regression during the NLI shared task. On a very challenging dataset containing TOEFL essays written by speakers of 11 different languages, TF-IDF with SVM reached 81.4% and 84.6% accuracy on the test set when using 10-fold cross validation.

Finally, for *Run 3* we use a simple, yet efficient and fast method that combines Laplace smoothing and a probabilistic classifier. The approach was previously applied to distinguish Brazilian and European Portuguese texts (Zampieri and Gebre, 2012) and it is available as an open source tool called *VarClass* (Zampieri and Gebre, 2014). The likelihood function is calculated as described in equation 1.

$$P(L|text) = \arg\max_{L} \sum_{i=1}^{N} \log P(n_i|L) + \log P(L) \quad (4)$$

where N is the number of n-grams in the test text, n_i is the ith n-gram and L stands for the language models. Given a test text, we calculate the probability for each of the language models. The language model with the highest probability determines the identified language of the text.

5 Results

We start by reporting the official shared task results in terms of accuracy. Table 2 highlights the best results for each dataset.

Run	Test Set A	Test Set B
Run 1	94.09%	**92.77%**
Run 2	**95.24%**	**92.77%**
Run 3	94.07%	92.47%
Rank	2nd *out of 9*	4th *out of 7*

Table 2: Overall accuracy.

Results obtained by the three systems are very similar. Nevertheless, the SVM with TF-IDF Weighting approach obtained slightly better overall performance (*Run 2*). As we expected, the systems' performance drops from test set A to test set B. This means that our systems rely on named entities to discriminate between similar languages. It is important to point out that we did not do any specific training with the blinded named entities. Probably we could have achieved better results if we had prepared our systems to cope with this variation.

Table 3 presents the accuracy obtained by our best system (SVM with TF-IDF Weighting - *Run 2*) for each of the 14 classes. The results show that our best system achieved perfect performance in two of the language groups (Czech/ Slovak and Bulgarian/ Macedonian), probably due to exclusive characters present in one of the languages, as well as in identifying the 'unseen' languages in test set A.

Language/Variety	Test Set A	Test Set B
Bosnian	83.5%	76.6%
Croatian	91.8%	92.2%
Serbian	93.9%	90.7%
Indonesian	99.2%	97.5%
Malay	99.4%	99.5%
Czech	100%	99.9%
Slovak	100%	100%
Brazilian Portuguese	93.6%	90.5%
European Portuguese	93.0%	86.7%
Argentine Spanish	91.2%	89.2%
Castilian Spanish	94.8%	94.5%
Macedonian	100%	100%
Bulgarian	100%	100%
Unknown	100%	99.8%

Table 3: *Run 2*: performance per language.

Although the performance did not drop for Croatian and Malay when comparing test set A and B as it did for the rest of the languages, we do not think that this reflects any property of Croatian nor Malay nor any characteristics of the dataset. This is a simple preference of the classifier when distinguishing Croatian from Bosnian and Serbian, and Malay from Indonesian.

Tables 4, 5 and 6 present the confusion matrices obtained by the three systems using the 2,000 gold test instances.

Table 6 shows that Likelihood Estimation used for *Run 3* achieved higher scores when discriminating between language varieties, by classifying 1,912 Peninsular Spanish texts and 1,867 Brazilian Portuguese texts correctly. On the other hand, it was the only method which did not score 100% when classifying 'unseen' languages. Due to its simplicity, this method is well suited to discriminate between language varieties, hence the good results obtained in binary classification for Portuguese (Zampieri and Gebre, 2012), but

	bg	bs	cz	es-AR	es-ES	hr	id	mk	my	pt-BR	pt-PT	sk	sr	xx
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1578	0	0	0	241	0	0	0	0	0	0	181	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1774	226	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	227	1773	0	0	0	0	0	0	0	0	0
hr	0	132	0	0	0	1841	0	0	0	0	0	0	26	1
id	0	0	0	0	0	0	1979	0	21	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	30	0	1970	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1826	174	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	222	1778	0	0	0
sk	0	0	0	0	0	0	0	0	0	0	0	2000	0	0
sr	0	86	0	0	0	41	0	0	0	0	0	0	1873	0
xx	0	0	0	0	0	0	0	0	0	0	0	0	0	2000

Table 4: Confusion Matrix *Run 1* - Axis Y represents the actual classes and Axis X the predicted classes.

	bg	bs	cz	es-AR	es-ES	hr	id	mk	my	pt-BR	pt-PT	sk	sr	xx
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1661	0	0	0	193	0	0	0	0	0	0	146	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1796	204	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	209	1791	0	0	0	0	0	0	0	0	0
hr	0	135	0	0	0	1843	0	0	0	0	0	0	21	1
id	0	0	0	0	0	0	1988	0	12	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	19	0	1981	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1844	156	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	166	1834	0	0	0
sk	0	0	1	0	0	0	0	0	0	0	0	1999	0	0
sr	0	86	0	0	0	41	0	0	0	0	0	0	1891	0
xx	0	0	0	0	0	0	0	0	0	0	0	0	0	2000

Table 5: Confusion Matrix *Run 2* - Axis Y represents the actual classes and Axis X the predicted classes.

	bg	bs	cz	es-AR	es-ES	hr	id	mk	my	pt-BR	pt-PT	sk	sr	xx
bg	2000	0	0	0	0	0	0	0	0	0	0	0	0	0
bs	0	1623	0	0	0	198	0	0	0	0	0	0	179	0
cz	0	0	2000	0	0	0	0	0	0	0	0	0	0	0
es-AR	0	0	0	1623	377	0	0	0	0	0	0	0	0	0
es-ES	0	0	0	88	1912	0	0	0	0	0	0	0	0	0
hr	0	205	0	0	0	1746	0	0	0	0	0	0	49	0
id	0	0	0	0	0	0	1980	0	20	0	0	0	0	0
mk	0	0	0	0	0	0	0	2000	0	0	0	0	0	0
my	0	0	0	0	0	0	8	0	1992	0	0	0	0	0
pt-BR	0	0	0	0	0	0	0	0	0	1867	133	0	0	0
pt-PT	0	0	0	0	0	0	0	0	0	236	1764	0	0	0
sk	0	0	0	0	0	0	0	0	0	0	0	2000	0	0
sr	0	107	0	0	0	36	0	0	0	0	0	0	1857	0
xx	5	2	0	5	7	3	0	0	0	0	0	0	2	1976

Table 6: Confusion Matrix *Run 3* - Axis Y represents the actual classes and Axis X the predicted classes.

it clearly does not cope well with unseen data. Consequently, this method can be considered a good choice for situations in which all classes are known *a priori*.

6 Conclusion

This paper presented the MMS entry to the Discriminating between Similar Languages (DSL) shared task. We submitted three different approaches to deal with the task in hand, and their overall scores turned out to be very similar. The linear SVM classifier combined with TF-IDF weighting (*Run 2*) achieved slightly better results than the other two methods, i.e. 95.24% against 94.07% and 94.09% accuracy on test set A. The system ranked 2nd (out of 9 teams) on the test set A and 4th (out of 7 teams) on the test set B.

Based on the results, we observed that the systems' performance drop from test set A to test set B. This was already expected because named entities play an important role in this kind of task. One of the ways to cope with the influence of named entities in text classification is to use delexicalised text representations relying on POS tags or hybrid representations mixing word forms and grammatical categories. In our previous work, however, the results obtained using POS tags to discriminate between Spanish varieties, indicate that the use of more abstract text representations do not result in performance gain (Zampieri et al., 2013). In future work we would like to return to the question of text representation and investigate whether we can propose features that deliver high performance across multiple datasets.

An interesting approach would be to model these three systems hierarchically. This would result in a two-level classification task, first identifying the language group (grouped by similarity) and then the language itself. This approach was proposed by the NRC team, the DSL winner of the 2014 edition (Goutte et al., 2014). In the future we plan to investigate whether performing classification on two levels would increase the overall score or not.

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n^{o} 317471.

References

Heba Elfardy and Mona T Diab. 2014. Sentence level dialect identification in Arabic. In *Proceedings of ACL*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskens. 2013. Improving native language identification with tf-idf weighting. In *Proceedings of the 8th BEA workshop*.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The NRC system for discriminating similar languages. In *Proceedings of the VarDial Workshop*.

Chu-ren Huang and Lung-hao Lee. 2008. Contrastive approach towards text source classification based on top-bag-of-word similarity. In *Proceedings of PACLIC*.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 137–142. Springer.

Ben King, Dragomir Radev, and Steven Abney. 2014. Experiments in sentence language identification with groups of similar languages. In *Proceedings of the VarDial Workshop*.

Nikola Ljubešić and Denis Kranjčić. 2015. Discriminating between closely related languages on twitter. *Informatica*, 39(1).

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Marco Lui and Paul Cook. 2013. Classifying English documents by national dialect. In *Proceedings of Australasian Language Technology Workshop*, pages 5–15.

Marco Lui, Ned Letcher, Oliver Adams, Long Duong, Paul Cook, and Timothy Baldwin. 2014. Exploring methods and resources for discriminating similar languages. In *Proceedings of VarDial*.

Wolfgang Maier and Carlos Gómez-Rodrıguez. 2014. Language variety identification in Spanish tweets. In *Proceedings of the LT4CloseLang Workshop*.

Shervin Malmasi and Mark Dras. 2015. Automatic Language Identification for Persian and Dari texts. In *Proceedings of PACLING 2015*, pages 59–64.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In *Proceedings of PACLING 2015*, pages 209–217, Bali, Indonesia, May.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Kavi Narayana Murthy and G Bharadwaja Kumar. 2006. Language identification from small text samples. *Journal of Quantitative Linguistics*, 13(01):57–80.

Vlad Niculae, Marcos Zampieri, Liviu P Dinu, and Alina Maria Ciobanu. 2014. Temporal text ranking and automatic dating of texts. In *Proceedings of EACL*. ACL.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Bali Ranaivo-Malançon. 2006. Automatic identification of close languages - case study: Malay and Indonesian. *ECTI Transactions on Computer and Information Technology*, 2:126–134.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of Arabic language varieties and dialects in social media. In *Proceedings of SocialNLP 2014*.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence level dialect identification for machine translation system selection. In *Proceedings of ACL*, pages 772–778, Baltimore, USA.

Gerard Salton and Michael J McGill. 1986. *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar, October.

Liling Tan, Marcos Zampieri, Nikola Ljubešić, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The DSL corpus collection. In *Proceedings of The Workshop on Building and Using Comparable Corpora (BUCC)*, Reykjavik, Iceland.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of BEA*, Atlanta, GA, USA, June.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.

Christoph Tillmann, Saab Mansour, and Yaser Al-Onaizan. 2014. Improved sentence-level Arabic dialect classification. In *Proceedings of the VarDial Workshop*, pages 110–119, Dublin, Ireland, August.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS*, pages 233–237.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2014. Varclass: An open source language identification tool for language varieties. In *Proceedings of Language Resources and Evaluation (LREC)*.

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2013. N-gram language models and POS distribution for the identification of Spanish varieties. In *Proceedings of TALN*, pages 580–587.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the VarDial Workshop*, pages 58–67.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the dsl shared task 2015. In *Proceedings of LT4VarDial*, Hissar, Bulgaria.

Arkaitz Zubiaga, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Vıctor Fresno. 2014. Overview of TweetLID: Tweet language identification at SEPLN 2014. In *Proceedings of SEPLN 2014*, pages 1–11, Girona, Spain.

Costa et al. (2015c)

Costa, H., Corpas Pastor, G., and Mitkov, R. (2015). **Measuring the Relatedness between Documents in Comparable Corpora**. In *11th Int. Conf. on Terminology and Artificial Intelligence, TIA'15*, pages 29-37, Granada, Spain.

Measuring the Relatedness between Documents in Comparable Corpora

Hernani Costa[a], **Gloria Corpas Pastor**[a] **and Ruslan Mitkov**[b]
[a]LEXYTRAD, University of Malaga, Spain
[b]RIILP, University of Wolverhampton, UK
{hercos,gcorpas}@uma.es, r.mitkov@wlv.ac.uk

Abstract

This paper aims at investigating the use of textual distributional similarity measures in the context of comparable corpora. We address the issue of measuring the relatedness between documents by extracting, measuring and ranking their common content. For this purpose, we designed and applied a methodology that exploits available natural language processing technology with statistical methods. Our findings showed that using a list of common entities and a simple, yet robust set of distributional similarity measures was enough to describe and assess the degree of relatedness between the documents. Moreover, our method has demonstrated high performance in the task of filtering out documents with a low level of relatedness. By a way of example, one of the measures got 100%, 100%, 95% and 90% precision when injected 5%, 10%, 15% and 20% of noise, respectively.

1 Introduction

Comparable corpora[1] can be considered an important resource for several research areas such as Natural Language Processing (NLP), terminology, language teaching, and automatic and assisted translation, amongst other related areas. Nevertheless, an inherent problem to those who deal with comparable corpora in a daily basis is the uncertainty about the data they are dealing with. Indeed, little work has been done on semi- or automatically characterising such linguistic resources and attempting a meaningful description of their content is often a perilous task (Corpas Pastor and Seghiri, 2009). Usually, a corpus is given a short description such as "casual speech transcripts" or "tourism specialised comparable corpus". Yet, such tags will be of little use to those users seeking for a representative and/or high quality domain-specific corpora. Apart from the usual description that comes along with the corpus, like number of documents, tokens, types, source(s), creation date, policies of usage, etc., nothing is said about how similar the documents are or how to retrieve the most related ones. As a result, most of the resources at our disposal are built and shared without deep analysis of their content, and those who use them blindly trust on the people's or research group's name behind their compilation process, without knowing nothing about the relatedness quality of the documents. Although some tasks require documents with a high degree of relatedness between each other, the literature is scarce on this matter.

Accordingly, this work explores this niche by taking advantage of several textual Distributional Similarity Measures (DSMs) presented in the literature. Firstly, we selected a specialised corpus about tourism and beauty domain that was manually compiled by researchers in the area of translation and interpreting studies. Then, we designed and applied a methodology that exploits available NLP technology with statistical methods to assess how the documents correlate with each other in the corpus. Our assumption is that the amount of information contained in a document can be evaluated via summing the amount of information contained in the member words. For

[1]I.e. corpora that include similar types of original texts in one or more language using the same design criteria (cf. (EAGLES, 1996; Corpas Pastor, 2001)).

this purpose, a list of common entities was used as a unit of measurement capable of identifying the amount of information shared between the documents. Our hypothesis is that this approach will allow us to: compute the relatedness between documents; describe and characterise the corpus itself; and to rank the documents by their degree of relatedness. In order to evaluate how the DSMs perform the task of ranking documents based on their similarity and filter out the unrelated ones, we introduced noisy documents, i.e. out-of-domain documents to the corpus in hand.

The remainder of the paper is structured as follows. Section 2 introduces some fundamental concepts related with DSMs, i.e. explains the theoretical foundations, related work and the DSMs exploited in this experiment. Then, Section 3 presents the corpora used in this work. After applying the methodology described in Section 4, Section 5 presents and discusses the obtained results in detail. Finally, Section 6 presents the final remarks and highlights our future work.

2 Distributional Similarity Measures

Information Retrieval (IR) (Singhal, 2001) is the task of locating specific information within a collection of documents or other natural language resources according to some request. This field is rich in statistical methods that use words and their (co-)occurrence to retrieve documents or sentences from large data sets. In simple words, these IR methods aim to find the most frequently used words and treat the rate of usage of each word in a given text as a quantitative attribute. Then, these words serve as features for a given statistical method. Following Harris' distributional hypothesis (Harris, 1970), which assumes that similar words tend to occur in similar contexts, these statistical methods are suitable, for instance to find similar sentences based on the words they contain (Costa et al., 2015) and automatically extract or validate semantic entities from corpora (Costa et al., 2010; Costa, 2010; Costa et al., 2011). To this end, it is assumed that the amount of information contained in a document could be evaluated by summing the amount of information contained in the document words. And, the amount of information conveyed by a word can be represented by means of the weight assigned to it (Salton and Buckley, 1988). Having this in mind, we took advantage of two IR measures commonly used in the literature, the Spearman's Rank Correlation Coefficient (SCC) and the Chi-Square (χ^2) to compute the similarity between documents written in the same language (see section 2.1 and 2.2). Both measures are particularly useful for this task because they are independent of text size (mostly because both use a list of the common entities), and they are language-independent.

The SCC distributional measure has been shown effective on determining similarity between sentences, documents and even on corpora of varying sizes (Kilgarriff, 2001; Costa et al., 2015; Costa, 2015). It is particularly useful, for instance to measure the textual similarity between documents because it is easy to compute and is independent of text size as it can directly compare ranked lists for large and small texts.

The χ^2 similarity measure has also shown its robustness and high performance. By way of example, χ^2 have been used to analyse the conversation component of the British National Corpus (Rayson et al., 1997), to compare both documents and corpora (Kilgarriff, 2001; Costa, 2015), and to identify topic related clusters in imperfect transcribed documents (Ibrahimov et al., 2002). It is a simple statistic measure that permits to assess if relationships between two variables in a sample are due to chance or the relationship is systematic.

Bearing this in mind, distributional similarity measures in general and SCC and χ^2 in particular have a wide range of applicabilities (Kilgarriff, 2001; Costa et al., 2015; Costa, 2015). Indeed, this work aims at proving that these simple, yet robust and high-performance measures allow to describe the relatedness between documents in specialised corpora and to rank them according to their similarity.

2.1 Spearman's Rank Correlation Coefficient (SCC)

In this work, the SCC is adopted and calculated as in Kilgarriff (2001). Firstly, a list of the common entities[2] L between two documents d_l and d_m is compiled, where $L_{d_l,d_m} \subseteq (d_l \cap d_m)$. It is possible to use the top n most common entities or all

[2] In this work, the term 'entity' refers to "single words", which can be a token, a lemma or a stem.

common entities between two documents, where n corresponds to the total number of common entities considered $|L|$, i.e. $\{n|n \in N^0, n \leq |L|\}$ – in this work we use all the common entities for each document pair, i.e. $n = |L|$. Then, for each document the list of common entities (e.g. L_{d_l} and L_{d_m}) is ranked by frequency in an ascending order ($R_{L_{d_l}}$ and $R_{L_{d_m}}$), where the entity with lowest frequency receives the numerical raking position 1 and the entity with highest frequency receives the numerical raking position n. Finally, for each common entity $\{e_1, ..., e_n\} \in L$, the difference in the rank orders for the entity in each document is computed, and then normalised as a sum of the square of these differences $\left(\sum_{i=1}^{n} s_i^2\right)$. The final SCC equation is presented in expression 1, where $\{SCC|SCC \in R, -1 \geq SCC \leq 1\}$.

$$SCC(d_l, d_m) = 1 - \frac{6 * \sum_{i=1}^{n} s_i^2}{n^3 - n} \quad (1)$$

2.2 Chi-Square (χ^2)

The Chi-square (χ^2) measure also uses a list of common entities (L). Similarly to SCC, it is also possible to use the top n most common entities or all common entities between two documents, and again, we use all the common entities for each document pair, i.e. $n = |L|$. The number of occurrences of a common entity in L that would be expected in each document is calculated from the frequency lists. If the size of the document d_l and d_m are N_l and N_m and the entity e_i has the following observed frequencies $O(e_i, d_l)$ and $O(e_i, d_m)$, then the expected values are $e_{i_{d_l}} = \frac{N_l * (O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$ and $e_{i_{d_m}} = \frac{N_m * (O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$. Equation 2 presents the χ^2 formula, where O is the observed frequency and E the expected frequency. The resulted χ^2 score should be interpreted as the interdocument distance between two documents. It is also important to mention that $\{\chi^2|\chi^2 \in R, 1 \geq \chi^2 < \infty\}$, which means that as more unrelated the common entities in L are, the lower the χ^2 score will be.

$$\chi^2(d_l, d_m) = \sum \frac{(O - E)^2}{E} \quad (2)$$

3 Corpora

INTELITERM[3] is a specialised comparable corpus composed of documents collected from the Internet. It was manually compiled by researchers with the purpose of building a representative corpus (Biber, 1988, p.246) for the Tourism and Beauty domain. It contains documents in four different languages (English, Spanish, Italian and German). Some of the texts are translations of each other (parallel), yet the majority is composed of original texts. The corpus is composed of several subcorpora, divided by the language and further for each language there are translated and original texts. For the purpose of this work, only original documents in English, Spanish and Italian were used, which for now on will be referred as int_en, int_es, int_it, respectively.

In order to analyse how the DSMs perform the task of ranking documents based on their similarity and filter out the unrelated ones, it is necessary to introduce noisy documents, i.e. out-of-domain documents to the various subcorpora. To do that, we chose the well-known Europarl[4] corpus (Koehn, 2005), a parallel corpus composed by proceedings of the European Parliament. As mentioned further in section 5.2, we added different amounts of noise to the various subcorpora, more precisely 5%, 10%, 15% and 20%. These noisy documents were randomly selected from the "one per day" Europarl v.7 for the three working languages: English, Spanish and Italian (eur_en, eur_es, eur_it, respectively).

	nDocs	types	tokens	$\frac{types}{tokens}$
int_en	151	11,6k	496,2k	0.023
eur_en	30	3.4k	29,8k	0.116
int_es	224	13,2k	207,3k	0.063
eur_es	44	5,6k	43,5k	0.129
int_it	150	19,9k	386,2k	0.052
eur_it	30	4,7k	29,6k	0.159

Table 1: Statistical information per subcorpora.

All the statistical information about both the INTELITERM subcorpora and the set of 20% of noisy documents, randomly selected for each working language, are presented in Table 1. In detail, this Table shows: the number of documents

[3] http://www.lexytrad.es/proyectos.html
[4] http://www.statmt.org/europarl/

(nDocs); the number of types (types); the number of tokens (tokens); and the ratio of types per tokens ($\frac{types}{tokens}$) per subcorpus. These values were obtained using the Antconc 3.4.3 (Anthony, 2014) software, a corpus analysis toolkit for concordancing and text analysis.

4 Methodology

This section describes the methodology employed to calculate and rank documents based on their similarity using Distributional Similarity Measures (DSMs). All the tools, libraries and frameworks used for the purpose in hand are also pointed out.

1) **Data Preprocessing**: firstly all the INTELITERM documents were processed with the OpenNLP[5] Sentence Detector and Tokeniser. Then, the annotation process was done with the TT4J[6] library, which is a Java wrapper around the popular TreeTagger (Schmid, 1995) – a tool specifically designed to annotate text with part-of-speech and lemma information. Regarding the stemming, we used the Porter stemmer algorithm provided by the Snowball[7] library. A method to remove punctuation and special characters within the words was also implemented. Finally, in order to get rid of the noise, a stopword list[8] was compiled to filter out the most frequent words in the corpus. Once a document is computed and the sentences are tokenised, lemmatised and stemmed, our system creates a new output file with all this new information, i.e. a new document containing: the original, the tokenised, the lemmatised and the stemmed text. Using the stopword list mentioned above a Boolean vector describing if the entity is a stopword or not is also added to the document. This way, the system will be able to use only the tokens, lemmas and stems that are not stopwords.

2) **Identifying the list of common entities between documents**: in order to identify a list of common entities (from now on we will use the acronym NCE), a co-occurrence matrix was built for each pair of documents. Only those that have at least one occurrence in both documents are considered. As required by the DSMs (see section 2), their frequency in both documents is also stored within this matrix ($L_{d_l,d_m} = \{e_i, (f(e_i,d_l), f(e_i,d_m)); e_j, (f(e_j,d_l), f(e_j,d_m)); ...; e_n, (f(e_n,d_l), f(e_n,d_m))\}$, where f represents the frequency of an entity in a document). With the purpose of analysing and comparing the performance of different DSMs, three different lists were created to be used as input features: the first one using the Number of Common Tokens (NCT), another using the Number of Common Lemmas (NCL) and the third one using the Number of Common Stems (NCS).

3) **Computing the similarity between documents**: the similarity between documents was calculated by applying three different DSMs ($DSMs = \{DSM_{NCE}, DSM_{SCC}, DSM_{\chi^2}\}$, where $_{NCE}$, $_{SCC}$ and $_{\chi^2}$ refer to Number of Common Entities, Spearman's Rank Correlation Coefficient and Chi-Square, respectively), each one calculated using three different input features (NCT, NCL and NCS).

4) **Computing the document final score**: the document final score $DSM(d_l)$ is the mean of the similarity scores of the document with all the documents in the collection of documents, i.e. $DSM(d_l) = \frac{\sum_{i=1}^{n-1} DSM_i(d_l,d_i)}{n-1}$, where n corresponds to the total number of documents in the collection and $DSM_i(d_l,d_i)$ the resulted similarity score between the document d_l with all the documents in the collection.

5) **Ranking documents**: finally, the documents were ranked in a descending order according to their DSMs scores (i.e. NCE, SCC or χ^2).

5 Results and Analysis

This experiment is divided into two parts. In the first part (section 5.1), we describe the corpus in hand by applying three different Distributional Similarity Measures (DSMs): the Number of Common Entities (NCE), the Spearman's Rank

[5] https://opennlp.apache.org
[6] http://reckart.github.io/tt4j/
[7] http://snowball.tartarus.org
[8] Freely available to download through the following URL https://github.com/hpcosta/stopwords.

Correlation Coefficient (SCC) and the Chi-Square (χ^2). As a input feature to the DSMs, three different lists of entities were used, i.e. the Number of Common Tokens (NCT), the Number of Common Lemmas (NCL) and the Number of Common Stems (NCS). By a way of example, Table 2 shows the NCT between documents, the SCC and the χ^2 scores and averages (av) along with the associated standard deviations (σ) per measure and subcorpus. Figure 1 presents the resulted average scores per document in a box plot format for all the combinations DSM vs. feature. Each box plot displays the full range of variation (from min to max), the likely range of variation (the interquartile range or IQR), the median, and the high maximums and low minimums (also know as outliers). It is important to mention that for the first part of this experiment (section 5.1) we did not use a sample, but instead the entire INTELITERM subcorpora in their original size and form, which means that all obtained results and made observations came from the entire population, in this case the English (int_en), Spanish (int_es) and Italian (int_it) subcorpora (for more details about the subcorpora see section 3). Regarding the second part of this experiment, we used the same subcorpora, but an additional percentage of documents was added to them in order to test how the DSMs perform the task of filtering out these noisy documents, i.e. out-of-domain documents (see 5.2). In detail, Figure 2 shows how the average scores decrease when injecting noisy documents and Table 3 presents how the DSMs performed when that noise was injected.

5.1 Describing the Corpus

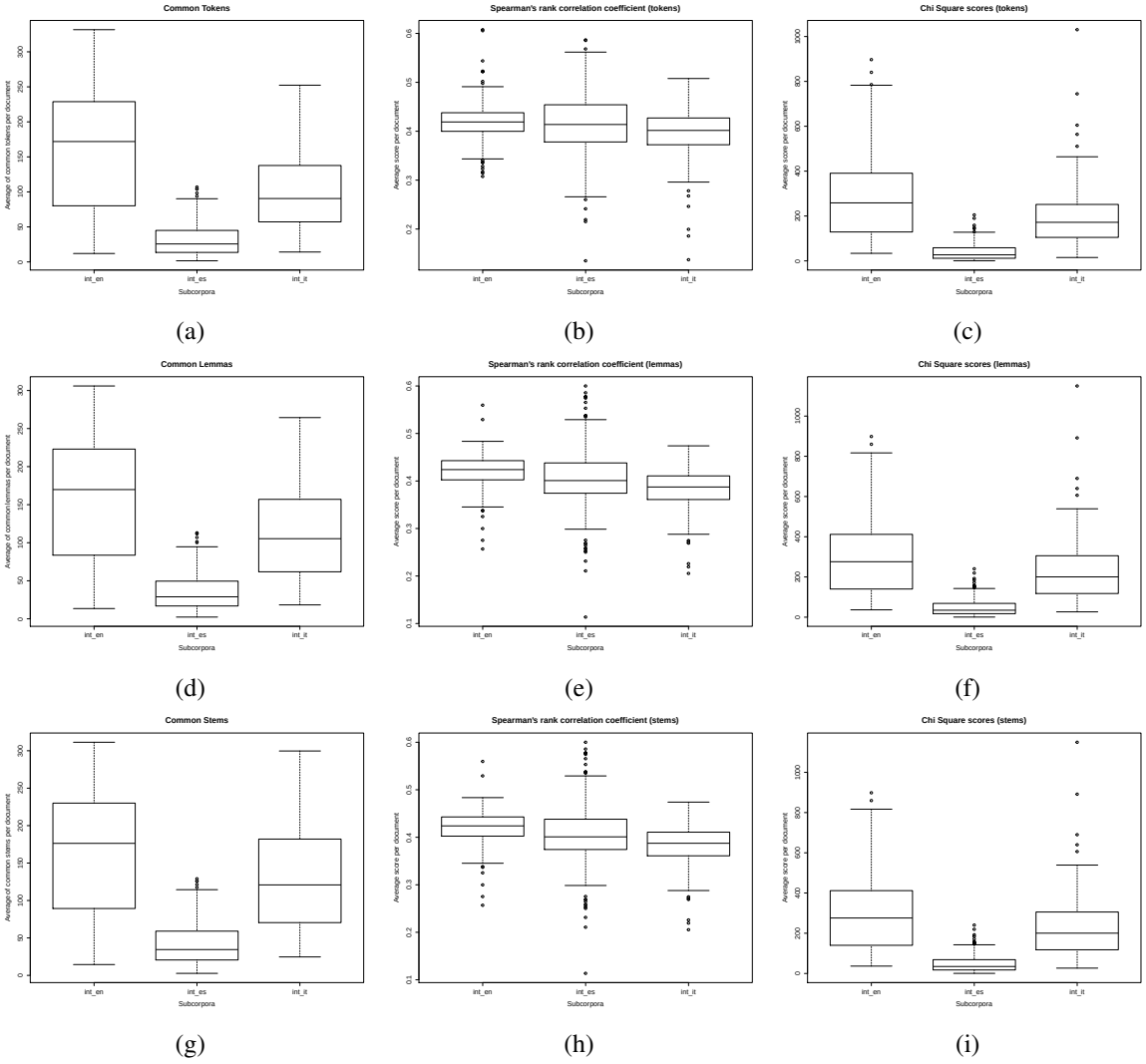
The first observation we can make from Figure 1 is that the distributions between the features are quite similar (see for instance Figures 1a, 1d and 1g). This means that it is possible to achieve acceptable results only using raw words (i.e. tokens). Stems and lemmas require more processing power and time to be used as features – especially lemmas due to the part-of-speech tagger dependency and time consuming process implied. In general, we can say that the scores for each subcorpus are symmetric (roughly the same on each side when cut down the middle), which means that the data is normally distributed. There are some exception that we will discuss along this section. Another interesting observation is related with the high Number of Common Tokens (NCT) in English (int_en) when compared with Italian and Spanish (int_it and int_es, respectively), see Table 2 and Figure 1a. Later in this section, we will try to explain this phenomenon.

SubC.	Stats	NCT	SCC	χ^2
int_en	av	163.70	0.42	279.39
	σ	83.87	0.05	177.45
int_es	av	31.97	0.41	40.92
	σ	23.48	0.07	38.21
int_it	av	101.08	0.39	201.97
	σ	55.71	0.05	144.68

Table 2: Average and standard deviation of common tokens scores between documents per subcorpus.

Although the NCT per document on average is higher for the int_en subcorpus, the interquartile range (IQR) is larger than for the other subcorpora (see Table 2 and Figure 1a), which means that the middle 50% of the data is more distributed and thus the average of NCT per document is more variable. Moreover, longest whiskers (the lines extending vertically from the box) in Figure 1a also indicates variability outside the upper and lower quartiles. Therefore, we can say that int_en has a wide type of documents and consequently some of them are only roughly correlated to the rest of the subcorpus. Nevertheless, the data is skewed left and the longest whisker outside the upper quartile indicates that the majority of the data is strongly similar, i.e. the documents have a high degree of relatedness between each other. This idea can be sustained not only by the positive average SCC scores, but also by the set of outliers above the upper whisker in Figure 1b. The average of 0.42 SCC score and σ=0.05 also implies a strong correlation between the documents in the int_en subcorpus (Table 2). Likewise, the longest whisker and the set of outliers outside the upper quartile in the χ^2 scores also indicate a high relatedness between the documents.

Regarding the int_it subcorpus, the SCC and the χ^2 scores (Figures 1b and 1c) and the average of 101.08 common tokens per document and σ=55.71 (Figure 1a and Table 2) suggest that the data is normally distributed (Figure 1b) and highly



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 1: INTELITERM: average scores between documents per subcorpus.

correlated. Although this subcorpus got lower average scores for all the DSMs when compared to the English subcorpus, Table 2, Figure 1a, 1b and 1c show that the average scores and the range of variation are quite similar to the English subcorpus. Therefore, we can conclude that the documents inside the Italian subcorpus are highly related between each other.

From the three subcorpora, the int_es subcorpus is the biggest one with 224 documents (Table 1). Nevertheless, the average scores per document are slightly different from the other box plots (see Figures 1a, 1b and 1c). The χ^2 standard deviation practically equal to its average (38.21 and 40.92, respectively) and the SCC variability inside and outside the IQR indicates some inconsistency in the data. Moreover, Table 2 and Figure 1a reveal a lower NCT compared with int_en and the int_it subcorpora.

The subcorpus int_en has 163 common tokens per document on average with a σ=83, and the subcorpora int_it and int_es only have 101 and 31 common tokens per document on average with a σ=55 and σ=23, respectively (Table 2, NCT column). This means that the int_it and int_es subcorpora are composed of documents with a lower level of relatedness when compared with the English one. This fact could happen because Italian and Spanish have a richer morphology compared to English. Therefore, due to bigger number of inflection forms per lemma, there is a larger number of tokens and consequently less common tokens per document in Spanish. Another explanation could come from the fact

that the tourism and beauty services are more developed in Italy and Spain than in the UK and therefore there are more variety on the vocabulary used as well as in the services offered. Indeed, Table 1 offers some evidences about the employed vocabulary. The English subcorpus has a lower number of types and a higher number of tokens (11,6k and 496,2k, respectively) when compared with the Italian (19,9k types and 386,2k tokens) and Spanish subcorpora (13,2k types and 207,3k tokens). The high difference on the average of common tokens per document between Spanish and the other two languages can also be related with the marketing strategies used to advertise tourism and beauty services, which is somehow hard to confirm. Despite that our method is able to catch the lexical level of similarity between the documents, the semantic level is not taken into account, i.e. does not consider synonyms as similar words for example, and consequently would result on slightly different similarity scores (again, another explanation difficult to confirm).

To conclude, we can state from the statistical and theoretical evidences that the int_en and the int_it subcorpora look like they assemble highly correlated documents. We can not say the same for the int_es subcorpus. Due to the scarceness of evidences, we can only not reject the idea that this subcorpus is composed of similar documents. Nevertheless, as we will see in the next section, the fact that int_es is composed by low related documents (according to our findings) will affect the ranking task.

5.2 Measuring DSMs Performance

The second part of this experiment aims at assessing how the DSMs perform the task of filtering out documents with a low level of relatedness. To do that, we injected different sets of out-of-domain documents, randomly selected from the Europarl corpus to the original INTELITERM subcorpora. More precisely, we injected 5%, 10%, 15% and 20%[9] to the various subcorpora. As we can see in Figure 2, the more noisy documents are injected, the lower is the NCT. Then, the methodology described in Section 4 was applied to these "new twelve subcorpora" (int_en05, int_en10, ..., int_it15 and int_it20, see Figure 2). As a result, at this point we have the documents ranked in a descending order according to their DSMs scores.

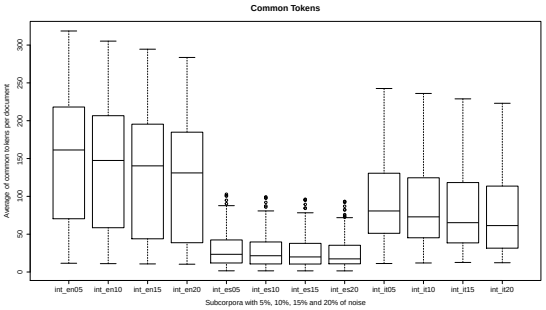


Figure 2: Average scores between documents when injecting 5%, 10%, 15% and 20% of noise to the various subcorpora.

In order to evaluate the DSMs precision, we analysed the first n positions in the ranking lists produced by the three DSMs (individually), and in this case n is the number of original documents in a given INTELITERM subcorpus. Table 3 presents the precision values obtained by the DSMs when injecting different amounts of noise to the various original subcorpora.

SubC	Noise	NCT	SCC	χ^2
int_en	*5%*	0.89	0.22	**1.00**
	10%	0.73	0.33	**1.00**
	15%	0.73	0.36	0.95
	20%	**0.80**	0.37	**0.90**
int_es	*5%*	0.00	0.00	0.38
	10%	0.07	0.07	0.20
	15%	0.09	0.09	0.17
	20%	0.14	0.18	0.23
int_it	*5%*	0.88	0.13	0.88
	10%	0.82	0.06	0.82
	15%	0.74	0.09	0.83
	20%	**0.73**	0.13	0.87

Table 3: DSMs precision when injecting different amounts of noise to the various subcorpora.

As expected, none of the DSMs got acceptable results for Spanish, being incapable of correctly identify noisy documents. However, we need to be aware that this happened due to the pre-existing low level of relatedness between the original documents in the int_es subcorpus (see Section 5.1 for more details). On the other hand, the DSMs show promising results for English and Italian. By

[9]The number of documents that correspond to these percentages can be inferred from Table 1.

a way of example, the χ^2 was capable of reaching 100% when injected 5% and 10% of noise to the int_en subcorpus, and even 90% when injected 20%. Although the NCT got lower precision, in general, when compared with the χ^2, it still reached 80% and 73% when injected 20% of noise to the English and to the Italian subcopora, respectively. From the evidences shown in Table 3, we can say that the NCT and the χ^2 are suitable for the task of filtering out low related documents with a high precision degree. The same cannot be say to the SCC measure, at least for this specific task.

6 Conclusions and Future Work

In this paper we presented a simple methodology and studied various Distributional Similarity Measures (DSMs) for the purpose of measuring the relatedness between documents in specialised comparable corpora. As input for these DSMs, we used three different input features (lists of common tokens, lemmas and stems). In the end, we conclude that for the data in hand these features had similar performance. In fact, our findings show that instead of using common lemmas or stems, which require external libraries, processing power and time, a simple list of common tokens was enough to describe our data. Moreover, we proved that it is possible to assess and describe comparable corpora through statistical methods. The number of entities shared by their documents, the average scores obtained with the SCC and the χ^2 measure resulted to be an important surgical toolbox to dissect and microscopically analyse comparable corpora.

Furthermore, these DSMs can be seen as a suitable tool to rank documents by their similarities. A handy feature to those who manually or semi-automatically compile corpora mined from the Internet and want to retrieve the most similar ones and filter out documents with a low level of relatedness. Our findings show promising results when filtering out noisy documents. Indeed, two of the measures got very high precision results, even when dealing with 20% of noise.

In the future, we intend not only to perform more experiments with these DSMs in other corpora and languages, but also test other DSMs, like Jaccard or Cosine and compare their performance.

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471. The research reported in this work has also been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. nº FFI2012-38881, 2012-2015); the R&D Project for Excellence TERMITUR (ref. nº HUM2754, 2014-2017); and the LATEST project (ref. 327197-FP7-PEOPLE-2012-IEF).

References

Laurence Anthony. 2014. AntConc (Version 3.4.3) Machintosh OS X. Waseda University. Tokyo, Japan. Available from http://www.laurenceanthony.net.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge, UK.

Gloria Corpas Pastor and Míriam Seghiri. 2009. Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). In A. Beeby, P.R. Inés, and P. Sánchez-Gijón, editors, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

Gloria Corpas Pastor. 2001. Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, 5(1):155–184.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2010. The Impact of Distributional Metrics in the Quality of Relational Triples. In *19th European Conf. on Artificial Intelligence, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, ECAI'10, pages 23–29, Lisbon, Portugal, August.

Hernani Costa, Hugo Gonçalo Oliveira, and Paulo Gomes. 2011. Using the Web to Validate Lexico-Semantic Relations. In *15th Portuguese Conf. on Artificial Intelligence*, volume 7026 of *EPIA'11*, pages 597–609, Lisbon, Portugal, October. Springer.

Hernani Costa, Hanna Béchara, Shiva Taslimipoor, Rohit Gupta, Constantin Orasan, Gloria Corpas Pastor, and Ruslan Mitkov. 2015.

MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. In 9^{th} *Int. Workshop on Semantic Evaluation*, SemEval'15, pages 96–101, Denver, Colorado, June. ACL.

Hernani Costa. 2010. Automatic Extraction and Validation of Lexical Ontologies from text. Master's thesis, University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering, Coimbra, Portugal, September.

Hernani Costa. 2015. Assessing Comparable Corpora through Distributional Similarity Measures. In *EXPERT Scientific and Technological Workshop*, pages 23–32, Malaga, Spain, June.

EAGLES. 1996. Preliminary Recommendations on Corpus Typology. Technical report, EAGLES Document EAG-TCWG-CTYP/P., May. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

Zelig Harris. 1970. Distributional Structure. In *Papers in Structural and Transformational Linguistics*, pages 775–794. D. Reidel Publishing Company, Dordrecht, Holland.

Oktay Ibrahimov, Ishwar Sethi, and Nevenka Dimitrova. 2002. The Performance Analysis of a Chi-square Similarity Measure for Topic Related Clustering of Noisy Transcripts. In 16^{th} *Int. Conf. on Pattern Recognition*, volume 4, pages 285–288. IEEE Computer Society.

Adam Kilgarriff. 2001. Comparing Corpora. *Int. Journal of Corpus Linguistics*, 6(1):97–133.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit*.

Paul Rayson, Geoffrey Leech, and Mary Hodges. 1997. Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. *Int. Journal of Corpus Linguistics*, 2(1):133–152.

Gerard Salton and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

Amit Singhal. 2001. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42.

Costa et al. (2015b)

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2015). **An Interpreters' Guide to Selecting Terminology Management Tools**. In *NATO Conf. on Terminology Management*, Brussels, Belgium.

An Interpreters' Guide to Selecting Terminology Management Tools

Hernani Costa, Gloria Corpas Pastor and Isabel Durá Muñoz
LEXYTRAD, University of Malaga, Spain
{hercos,gcorpas,iduran}@uma.es

Abstract

Time is golden... especially in the case of interpreters. Prior to a service, it is essential to search for relevant information and domain-specific terminology within a very limited period of time. During the interpreting service the information gathered needs to be easily accessible at all times. Similarly, after a given service, interpreters should be ideally able to store terms and any other documentation for future reference. In this context, choosing the right tool for a specific project can have a significant impact on the amount of time required to extract, manage and consult terminology before, during and after the interpreting service. Saving time from searches and management could have a positive impact on the overall quality of interpreting. This paper focuses on terminology management tools. We offer a set of measurable features that can be used to guide interpreters when choosing the most adequate terminology management tool for a given interpretation project. Then, we present the better-classified tools based on our findings. And finally, we briefly describe three semi-automatically terminology extraction tools that can be used during the preparation stage to identify relevant terms from text.

Extended Abstract

Interpreters often work in a wide range of domains and have limited time to prepare themselves for a given interpreting service. To ensure the best possible results during the interpretation process, interpreters usually perform an extensive search for specialised knowledge and terminology as they need to familiarise themselves with concepts, technical terms, and proper names in the interpreters' working languages. Moreover, especially in consecutive interpreting and in a booth, they rely on these findings to help them during the interpretation process. Unlike translators, for whom computer-assisted tools make part of their translation pipeline for several years already, interpreters have not benefited from the same level of innovation. We can even say that their work relies by and large on traditional or manual methods. Fortunately, there are currently several terminology extraction and management tools capable of assisting interpreters before and during an interpretation service. Our communication aims not only to show how interpreters can benefit from these technology tools in their daily work but also how to evaluate them. In detail, we intend to demonstrate that it is possible to create a set of measurable features that can be used to access and distinguish the different Terminology Management Tools (TMT) available on the market and consequently ensure the choice of the best tool for a given interpretation project. Apart from that, we mention the most complete TMTs based on our findings. And finally, we briefly describe three semi-automatically Terminology Extraction Tools (TET) that can be used to identify relevant from text during the preparation stage.

As we know TMS differ from one another in their functionalities, practical issues, degrees of user-friendliness and target audience (i.e. individual or enterprise usage). Therefore, it is necessary to establish a set of specific and measurable features that permit us to assess and distinguish the different tools concerning individual's and company's needs in such a way that the results would be useful for both potential customers as well as to the designers of such systems. Departing from the conclusions drawn from the literature review (cf. Bilgen (2011); Rodríguez and Schnell (2009); Costa et al. (2014a) and Costa et al. (2014b)) and a careful analysis of the priorities for the design and features to be included

in a TMT, we identified 15 measurable features. For instance, the "freedom to define the basic structure" identified by Rodríguez and Schnell (2009) was reformulated into several practical measurable features, such as "Nº of descriptive fields", "Nº of working languages" and "Nº of languages per glossary". Moreover, the possibility of "developing multilingual mini-databases", also identified in their study, was reconsidered as measurable features by means of the following criteria: "Manages multiple glossaries" and "Nº of languages per glossary". Another example is the "Remote Glossary Exchange" measurable feature, which was inferred from the study conducted by Bilgen (2011), who identified the need to exchange terminological information. For more details about these features see Costa et al. (2014b). Based on this comparative analysis, none of the investigated TMTs exhibit all the desirable features. Nevertheless, SDL MultiTerm was the best classified standalone TMT with 77 points out of 100. Another interesting finding in our research was that web-based TMTs are more useful to share terminology and all the 6 web-based TMS that we analysed got similar scores, ranging from 74 (Acrolinx) to 78 (flahterm). Despite mobile TMS do not get acceptable scores when compared with standalone and web-based TMTs – Glossary Assistant got 53 and The Interpreter's Wizard 39 points – and they do not offer the necessary comfort to manage terminology, they still play an important role when a quick search for terminology is required, e.g. while in a booth. Although TETs are not totally accurate when used to semi-automatically extract terminology, they are the faster option available to identify for example the most frequent words or lexical units. For example, TermSuite (Daille (2012)) is an open-source and platform-independent TET that allows to extract bilingual terminology from comparable corpora in five European and two non-European languages. Also using statistic-based methods, Rainbow and ExtPhrJ are two examples of open-source platform-independent TETs that can be freely used to extract terms, from monolingual text, in almost any language.

To conclude, our main findings suggest that most TMT are not envisaged to be used by interpreters. Therefore, TMT do not fulfil completely their needs and technology-assisted interpreting tools still have a long way to go when compared with computer-assisted tools for translators. In the future we intend to identify the most relevant features that a TET should have in order to help interpreters before the interpretation service.

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement nº 317471. Also, the research reported in this work has been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. nº FFI2012-38881, 2012-2015), and the R&D Project for Excelence TERMITUR (ref. nº HUM2754, 2014-2017).

References

Bilgen, B. (2011). *Investigating Terminology Management for Conference Interpreters: A User-oriented Study*. LAP Lambert Academic Publishing.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014a). A comparative User Evaluation of Terminology Management Tools for Interpreters. In *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, pages 68–76, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014b). Technology-assisted Interpreting. *MultiLingual 143*, 25(3):27–32.

Daille, B. (2012). Building Bilingual Terminologies from Comparable Corpora: The TTC TermSuite. In *5th Workshop on Building and Using Comparable Corpora (BUCC'12)*, pages 29–32, Istanbul, Turkey.

Rodríguez, N. and Schnell, B. (2009). A Look at Terminology Adapted to the Requirements of Interpretation. *Language Update*, 6(1):21–27.

Costa et al. (2015e)

Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015). **Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora**. In *New Horizons in Translation and Interpreting Studies (Full papers)*, pages 133-141, Geneva, Switzerland. Tradulex.

New Horizons in Translation and Interpreting Studies (Full papers) • December 2015 • pp.133-141

Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora

Hernani Costa[a], Gloria Corpas Pastor[a],
Ruslan Mitkov[b] and Miriam Seghiri[a]

[a]LEXYTRAD, University of Malaga, Spain
[b]RIILP, University of Wolverhampton, UK
{hercos,gcorpas}@uma.es, r.mitkov@wlv.ac.uk, seghiri@uma.es

Abstract

This article presents an ongoing project that which aims to design and develop a robust and agile web-based application capable of semi-automatically compiling multilingual comparable and parallel corpora, named iCorpora. Its main purpose is to increase the flexibility and robustness of the compilation, management and exploration of both comparable and parallel corpora. iCorpora intends to fulfil not only translators' and interpreters' needs, but also the needs of other professionals and laypeople, either by solving some of the usability problems found in the current compilation tools available on the market or by reducing their limitations and performance issues.

1 Introduction

In the last decade, there has been a growing interest in bilingual and multilingual corpora. In translation, in particular, their benefits have been demonstrated by several authors (cf. Bowker and Pearson, 2002; Bowker, 2002; Zanettin et al., 2003; Corpas Pastor, 2008; Corpas Pastor and Seghiri, 2009). Their objectivity, reusability, multiplicity and applicability of uses, easy handling and quick access to large volumes of data are just some of their advantages. Thus, it is not surprising that the use of corpora has been considered an essential resource in several research domains such as translation, terminology, language teaching, and automatic and assisted translation, amongst others. In particular, parallel corpora have become a very important source of knowledge, especially for Machine Translation (MT). Example-Based Machine Translation (EBMT) and Statistical Machine Translation (SMT) are just some examples of MT sub-areas where this kind of resource is fundamental, e.g. for the process of training (Hutchins and Somers, 1992). Nevertheless, the lack of sufficient/up-to-date parallel corpora and linguistic resources for narrow domains and poorly-resourced languages is currently one of the major obstacles to further advancement in these areas. One potential solution to the insufficient parallel translation data is the exploitation of non-parallel bilingual and multilingual text resources, also known as comparable corpora – i.e. corpora that include similar types of original texts in one or more language using the same design criteria (cf. EAGLES, 1996; Corpas Pastor, 2001:158). Although comparable corpora can compensate for the shortage of linguistic resources and ultimately improve automated translation quality for under-resourced languages and narrow domains, the problem of data collection is still a significant technical challenge.

Bearing this in mind, the iCorpora project (cf. Costa et al., 2014c; 2015) proposes not

New Horizons in Translation and Interpreting Studies (Full papers) • December 2015 • pp.133-141

only to create a user-friendly interface to compile parallel corpora, but also to exploit comparable corpora from the Web. Broadly speaking, this ambitious project aims to increase the flexibility and robustness of the compilation, management and exploration of both comparable and parallel corpora by creating a new web-based application from scratch.

2 Existing Corpora Compilation Tools

The World Wide Web has become a primary meeting place for information and recreation, for communication and commerce. Millions of users have created billions of webpages in which they expressed their views about the world. As a source of machine-readable texts for corpus linguists and researchers in related fields such as Natural Language Processing (NLP) and MT for example, the Web offers extraordinary accessibility, quantity, variety and cost-effectiveness. To this end, several tools (e.g. web crawlers, language identifiers, HTML parsers, HTML cleaners, etc.) have been developed and combined in order to produce corpora from this 'goldmine'. Therefore, this section aims to describe the most relevant approaches, methodologies, and tools capable of exploiting parallel and comparable corpora from the Web.

2.1 Mining Parallel Corpora

The Internet can be already considered a large multilingual corpus due to its huge number of multilingual websites, in which different pages can contain the same written text in different languages. This means that some of their webpages can be paired into *bitexts* (or parallel texts) – a very important source of knowledge, especially for MT systems. Nevertheless, the problem of collecting these data is still a significant technical challenge and the question remains: How can we find these parallel texts and obtain an aligned parallel corpus from them? Some attempts to answer this question are presented below.

STRAND[1] (Structural Translation Recognition, Acquiring Natural Data) (Resnik, 1998; 1999; Resnik and Smith, 2003) can be considered as one of the earliest core web-mining architectures capable of identifying webpages which are candidates to be bitexts. In order to do this, it uses the structural features of documents, a content-based measure of translational equivalence, and the Web as a source for mining bitexts on a large scale. The general procedure includes three main steps: 1) locate possibly parallel webpages; 2) generate candidate pairs of parallel webpages; and, finally, 3) apply structural filters to the candidate set. The details about the process can be found in Resnik, 1998; 1999; Resnik and Smith, 2003.

Bitextor[2,3] (Esplà Gomis, 2009; Esplà Gomis and Forcada, 2009; 2010) is a free/open-source application created for Unix platforms, which aims to generate translation memories using multilingual websites as a corpus source. This tool was created to be as adaptable as possible when retrieving multilingual data from any kind of website and work with any pairs of languages. To do that, it combines context-based and URL-based heuristics to harvest aligned *bitexts* from multilingual websites. The Bitextor workflow can be divided into three main steps: 1) downloading, processing and choosing the parameters for the comparison; 2) webpage comparison; and, finally, 3) aligning the obtained webpages. It is important to mention that Bitextor is based on two main assumptions: parallel pages should be under the same domain and they should have similar HTML structure.

Although this section only describes two systems, BITS (Ma and Liberman,

[1] http://www.umiacs.umd.edu/~resnik/strand/
[2] http://bitextor.sourceforge.net
[3] http://sourceforge.net/projects/bitextor

New Horizons in Translation and Interpreting Studies (Full papers) • December 2015 • pp.133-141

1999), PTMiner (Chen and Nie, 2000), WeBiTex[4] (Désilets et al., 2008) and ILSP-FC (Papavassiliou et al., 2013) should also be mentioned as they were developed for the same purposes.

2.2 Mining Comparable Corpora

There is a growing literature on using the Web for constructing various types of text collections, including domain-specific monolingual, bilingual and multilingual comparable corpora. Although the process of compiling comparable corpora can be manually performed, nowadays specialised tools can be used to automate this tedious task. This section presents the two best-known tools on the market for exploiting corpora mined from the Web.

BooTCaT[5] (Baroni and Bernardini, 2004) is a free and open-source semi-automatic compilation application that makes use of online information to construct web-based corpora. The process is very simple and only requires a set of seed terms as input. Then, these seeds are randomly grouped to form tuples (i.e. a variety of combinations of the seeds), which are submitted as search query strings to a search engine. It is possible to build a larger corpus by repeating the process using more seeds, or even create a comparable corpus by repeating the process using translational equivalents. Despite the multiple advantages, BootCaT has a few limitations, which restricts the "natural process" that is usually used to compile bilingual or multilingual comparable corpora (cf. Baroni and Bernardini, 2004:1313 and Gutiérrez Florido et al., 2013:3).

WebBootCat[6] (Baroni et al., 2006) is similar to BootCaT, but instead of having to download and install the application, WebBootCat can be used online. Yet, it is only freely available on a trial basis or through subscription.

Although designed for other purposes, Terminus[7] and Corpográfo[8] should also be mentioned as examples of web-based compilation tools.

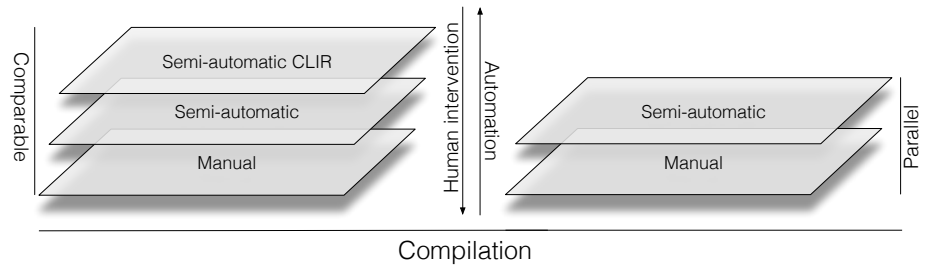
3 iCorpora: Compiling, Managing and Exploring Multilingual Data

As shown in the previous section, several semi-automatic compilation tools have been proposed so far, capable of exploiting either comparable or parallel corpora from the Web. However, these compilation tools are sometimes scarce, proprietary, simplistic with limited features or too complex to be used by laypeople. Moreover, comparable compilation tools were built to compile one monolingual corpus at a time and do not cover the entire compilation process (i.e. apart from compiling monolingual comparable corpora, they do not allow the managing and exploration of both parallel and multilingual comparable corpora). Thus, their simplicity, lack of features, performance issues and usability problems result in a pressing need to design new compilation tools tailored to fulfil not only translators' and interpreters' needs (cf. Costa et al. (2014b;a)), but also the needs of professionals and laypeople.

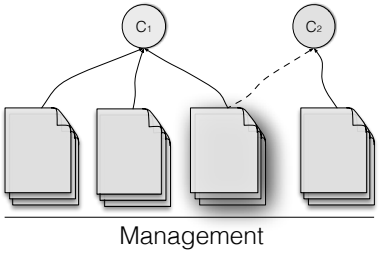
After a careful analysis of the shortcomings and strengths of the current compilation tools, we started designing and developing a robust and agile web-based application prototype to semi-automatically compile, manage and explore both parallel and multilingual comparable corpora, which we named *iCorpora*. In detail, iCorpora will aggregate three applications: *iCompileCorpora*, *iManageCorpora* and *iExploreCorpora*.

[4] http://www.webitext.com
[5] http://bootcat.sslmit.unibo.it
[6] https://www.sketchengine.co.uk/documentation/wiki/Website/Features#WebBootCat
[7] http://terminus.iula.upf.edu//cgi-bin/terminus2.0/terminus.pl
[8] http://www.linguateca.pt/corpografo/

New Horizons in Translation and Interpreting Studies (Full papers) • December 2015 • pp.133-141



(a) *iCompileCorpora.*



(b) *iManageCorpora.*

(c) *iExploreCorpora.*

Figure 1: *iCorpora layered model.*

3.1 iCompileCorpora

iCompileCorpora can be simply described as a web graphical interface which will guide the user through the entire corpus compilation process. It will not only provide a simple interface with easy-to-follow steps, but will also enable experienced users to set advanced compilation options during the process.

3.1.1 Compiling Comparable Corpora

The dimensions that comprise iCompileCorpora can be represented in a layered model comprising a manual, a semi-automatic web-based and a semi-automatic Cross-Language Information Retrieval (CLIR) layer (Figure 1a). This design option will not only result in increase of the flexibility and robustness of the compilation process, but will also hierarchically extend the manual layer features to the semi-automatic web-based layer and then to the semi-automatic CLIR layer. Specifically, the manual layer represents the option of compiling monolingual and multilingual comparable corpora, and will enable the manual upload of documents from a local or remote directory onto the platform. The second layer will permit the exploitation of both mono- and multilingual comparable corpora mined from the Internet. Although this layer can be considered similar to the approaches used by BootCaT and WebBootCat (see section 2.2), it has been designed to address some of their limitations (e.g. by allowing the use of more than one Boolean operator when creating search query strings). As there is now an increasing demand for systems that can somehow cross the language boundaries by retrieving information in various languages with just one query, the third layer aims to meet this demand by taking advantage of CLIR techniques to find relevant information written in a language different to the one semi-automatically retrieved by the methodology used in the previous layer.

3.1.2 Compiling Parallel Corpora

Regarding the parallel compilation process, iCompileCorpora will also facilitate for the manual upload of parallel documents from a local or remote directory onto the platform

New Horizons in Translation and Interpreting Studies (Full papers) • December 2015 • pp.133-141

(Figure 1a, manual layer). The second layer, i.e. the semi-automatic layer will offer the option of exploring parallel corpora mined from the Web. As shown in section 2.1, acquiring parallel data involves several tasks, such as crawling the web, parsing the structure of each fetched webpage and extracting its metadata, cleaning, classifying text, identifying near-duplicates, etc. Bearing this in mind, efficient focused web crawlers can be built by adapting existing open-source frameworks like Heritrix[9], Nutch[10] and Bixo[11]. Search engine Application Programming Interfaces (APIs) can also be used to identify in-domain webpages (Hong et al., 2010) or multilingual web sites (Resnik and Smith, 2003). At this point it is not yet clear which approach/algorithms and/or frameworks iCompileCorpora will use. Nevertheless, the methodology proposed in Resnik, 1998; 1999; Resnik and Smith, 2003 seems to be the most commonly used, i.e. locate possibly parallel webpages, generate candidates pairs of parallel webpages, and then apply structural filters to the candidate set in order to clean "noisy data".

3.2 iManageCorpora

The second application is called iManageCorpora (Figure 1b). This application will be specially designed to: manage (i.e. make it possible to edit, copy and paste sentences and documents from and to documents and corpora respectively, as well as to manage corpora into domains and sub-domains); measure the similarity between documents; and explore the representativeness of the corpora (cf. Corpas Pastor and Seghiri, 2009).

3.3 iExploreCorpora

Finally, iExploreCorpora (Fig. 1c) intends to offer a set of concordance features, such as the ability to search for words in context and automatically extract the most frequent words and multiword units, amongst other features.

[9] http://crawler.archive.org/
[10] http://nutch.apache.org
[11] http://openbixo.org/

4 Concluding Remarks

Against the background of the increasing importance of multilingual data, iCorpora's objectives are to develop a novel, flexible and robust web-based application for the compilation, management and exploitation of comparable and parallel corpora and to address the needs of translators and interpreters as well as other professional and casual users. This ongoing project aims to increase the flexibility and robustness of the compilation process by solving some of the usability problems found in the current compilation tools available on the market or by reducing their limitations and performance issues. By the end of this project, we intend to make this compilation tool publicly available, both in a research and in a commercial setting.

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n 317471. The research reported in this work has also been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. n FFI2012-38881, 2012-2015); the R&D Project for Excellence TERMITUR (ref. n HUM2754, 2014-2017); and the LATEST project (ref. 327197-FP7-PEOPLE-2012-IEF). We would also like to thank Emma Franklin for proof-reading this paper.

New Horizons in Translation and Interpreting Studies (Full papers) • December 2015 • pp.133-141

References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *4th Int. Conf. on Language Resources and Evaluation*, LREC'04, pages 1313–1316.

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *11th Annual Conf. of the European Association for Machine Translation*, EAMT'06, pages 247–252, Oslo, Norway. The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway).

Bowker, L. (2002). *Computer-aided Translation Technology: A Practical Introduction*. Didactics of translation series. University of Ottawa Press.

Bowker, L. and Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Chen, J. and Nie, J.-Y. (2000). Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. In *6th Conf. on Applied Natural Language Processing*, pages 21–28.

Corpas Pastor, G. (2001). Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, 5(1):155–184.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Studien Zur Romanischen Sprachwissenschaft Und Interkulturellen Kommunikation, 49. Peter Lang Pub Incorporated, Frankfurt, Germany.

Corpas Pastor, G. and Seghiri, M. (2009). Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). In Beeby, A., Inés, P., and Sánchez-Gijón, P., editors, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014a). A comparative User Evaluation of Terminology Management Tools for Interpreters. In *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, pages 68–76, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014b). Technology-assisted Interpreting. *MultiLingual #143*, 25(3):27–32.

Costa, H., Corpas Pastor, G., and Seghiri, M. (2014c). iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, London, UK.

Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015). iCorpora: Compiling, Managing and Exploring Multilingual Data. In *7th Int. Conf. of the Iberian Association of Translation and Interpreting Studies*, AIETI, pages 74–76, Malaga, Spain.

Désilets, A., Farley, B., Stojanovic, M., and Patenaude, G. (2008). WeBiText: Building Large Heterogeneous Translation Memories from Parallel Web Content. In *Translating and the Computer 30*, London, UK.

EAGLES (1996). Preliminary Recommendations on Corpus Typology. Technical report, EAGLES Document EAG-TCWG-CTYP/P. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

Esplà Gomis, M. (2009). Bitextor, un cosechador automático de memorias de traducción a partir de sitios web multilingües. *Procesamiento del Lenguaje Natural*, 43(1):365–366.

New Horizons in Translation and Interpreting Studies (Full papers) • December 2015 • pp.133-141

Esplà Gomis, M. and Forcada, M. (2009). Bitextor, a free/open-source software to harvest translation memories from multilingual websites. In *Workshop Beyond Translation Memories: New Tools for Translators MT*, Ottawa, Ontario, Canada.

Esplà Gomis, M. and Forcada, M. (2010). Combining Content-Based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93(1):77–86.

Gutiérrez Florido, R., Corpas Pastor, G., and Seghiri, M. (2013). Using semi-automatic compiled corpora for medical terminology and vocabulary building in the healthcare domain. In *10th Int. Conf. on Terminology and Artificial Intelligence (TIA'13), Workshop on Optimizing Understanding in Multilingual Hospital Encounters*, Paris, France.

Hong, G., Li, C.-H., Zhou, M., and Rim, H.-C. (2010). An Empirical Study on Web Mining of Parallel Data. In *23rd Int. Conf. on Computational Linguistics*, COLING'10, pages 474–482. ACL.

Hutchins, W. J. and Somers, H. L. (1992). *An Introduction to Machine Translation*. Academic Press.

Ma, X. and Liberman, M. (1999). BITS: A Method for Bilingual Text Search over the Web. In *Machine Translation Summit VII*.

Papavassiliou, V., Prokopidis, P., and Thurmair, G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *6th Workshop on Building and Using Comparable Corpora (BUCC'13)*, pages 43–51, Sofia, Bulgaria. ACL.

Resnik, P. (1998). Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text. In *3rd Conf. of the Association for Machine Translation in the Americas*, AMTA'98, Langhorne, PA, USA. Lecture Notes in Artificial Intelligence 1529.

Resnik, P. (1999). Mining the Web for Bilingual Text. In *37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL'99, pages 527–534. ACL.

Resnik, P. and Smith, N. (2003). The Web as a Parallel Corpus. *Computational Linguistics*, 29(3):349–380.

Zanettin, F., Bernardini, S., and Stewart, D. (2003). *Corpora in Translator Education*. Manchester: St. Jerome Publishing.

Costa et al. (2016b)

Costa, H., Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2016). **Nine terminology extraction Tools: Are they useful for translators?** *MultiLingual* #159, 27(3).

Nine Terminology Extraction Tools: Are they useful for translators?

Hernani Costa, Anna Zaretskaya, Gloria Corpas Pastor, Miriam Seghiri
University of Malaga
Malaga, Spain
{hercos,annazar,g.corpas,seghiri}@uma.es

Abstract

Terminology extraction tools have become an indispensable resource in education, research and business. Today, users can find a great variety of terminology extraction tools of all kinds, and they all offer different features. Apart from many other areas, these tools are especially helpful in the professional translation setting. We do not know, however, if the existing tools have all the necessary features for this kind of work. In search for the answer, we make an overview of nine selected tools available on the market and find out if they provide the translators' most favourite features.

1 Terminology extraction tools and their areas of application

The purpose of terminology extraction tools (TET) is to help users build terminological resources in a (semi-)automatic way. The need for such resources comes mostly from the growing needs in information management and translation, which make it more and more necessary to have some automated assistance when performing terminology-related tasks. Companies, freelancers and professionals in various linguistic fields can resort to these tools to, for example build glossaries, thesauri and terminological dictionaries that they use directly in their work. Moreover, TE is embedded in a number of natural language processing and linguistic research tasks, such as automatic indexing, machine translation, information extraction, creation of ontologies and knowledge bases, and corpus analysis. Although they have such broad range of applications, these tools are often designed for one specific purpose, which consequently makes their usage challenging when employed in a different setting.

One of the most important areas where terminology extraction is extremely helpful is in the translation industry. Today, more and more language service providers (LSP) as well as freelance translators and interpreters understand the benefits of automatizing terminology tasks. It not only allows them to quickly identify the domain of the documents they are dealing with, but also to easily find words and phrases that need to be paid special attention to. While translating terminological units, in many cases it is necessary to consider the domain and look up the term equivalents in special resources like terminology databases. And in addition, it helps maintain terminological consistency throughout the project between all the parts involved: the translator, the LSP and the client.

Apart from saving time, another significant advantage of using TET instead of manual terminology search consists in the possibility to specify different search criteria, which allows to adapt the search query to a particular task. This allows users to see all kinds of information they need about the term, and also to narrow the search and filter the results depending on what they are looking for. As an example, many state-of-the-art TET offer a possibility to see linguistic and statistic information about the term, the context where it appears, specify the number of words in the term, and many other useful features. Unfortunately, not every TET offers a full set of desirable features and settings, which makes it sometimes challenging to find the perfect tool for

the task in hand. Apart from the functionalities they offer, TET also differ as to the environment they work in. For instance, standalone installable tools require an installation process and work as independent computer programs. There also exist web-based tools, which work within a browser. And finally, there are reusable software that facilitates the development of larger applications, called frameworks.

Considering the existing variety, it is not clear how a professional translator is to proceed when choosing a TET suitable for the job. As we will see further, there are some TET that are specifically created for translators. But do they have all the necessary characteristics for translators? And, furthermore, what exactly are these characteristics?

2 Standalone Terminology Extraction Tools

Standalone software is probably the most popular type of software today, and TET are no exception. Standalone TET are tools that can be installed on the computer and operate independently of any other device or system.

SDL MultiTerm Extract is one of such applications. It is a component of SDL MultiTerm, a commercial terminology management tool that provides one solution to store, extract and manage multilingual terminology. Multiterm exists as a standalone application, and can also be integrated in SDL Trados Studio. It is one of the few tools that were designed specifically to be used by translators and is probably the most well-known TET in the translation industry. This TE system locates potential monolingual and bilingual terminology in documents and translation memories using a statistic-based method. The user can validate the extracted candidate terms by looking at a monolingual or bilingual concordance. A big advantage of this tool is its support for any language, including Unicode languages. In addition, it offers a number of functionalities that are useful in different translation scenarios, such as ability to compile a dictionary from parallel texts; flexible filtering that ensures that only the most frequent candidate terms are extracted; possibility to store an unlimited number of terms in any language; import and export glossaries from and to different technology environments. In addition, its integration with SDL Multiterm gives access to many convenient term-management functions, such as manually adding a variety of meta-data information to the terms, such as synonyms, context, definitions, illustrations, part-of-speech tags, URLs, etc., and searching not only the indexed terms but also their descriptive fields.

Simple Extractor as its name implies, offers significantly less functionalities compared to the previous tool. It is a commercial TET developed by DAIL Software S.L. for Mac OS, Linux and Windows platforms. This clean and easy-to-use standalone Java application was designed to automatically extract the most frequent words and multi-word terms from English, Portuguese, Spanish, French and Russian documents. Simple Extractor not only permits to extract a list of terms (from unigrams up to seven-grams), but also specify the minimum and maximum number of occurrences of a term. Moreover, Simple Extractor offers an option to load stopword lists, an advanced search functionality that permits to search through the extracted list of terms, to explore all the contexts that a specific term appears, to edit the term text, to filter the extracted terms according to the number of words that form them, and to sort the displayed output by any of its fields (frequency, term and context in alphabetical order). Finally, Simple Extractor permits to print out or export to a file (.pdf, .doc, .csv or .txt) all the extract terms, as well as their frequencies and corresponding contexts.

TermSuite is an open-source and platform-independent TET written in Java and distributed under the Apache License 2.0. It was developed within the scope of the TTC (Terminology Extraction, Translation Tools and Comparable Corpora) project, whose purpose was to design a tool capable of extracting bilingual terminology from comparable corpora in seven languages: English, French, German, Spanish, Chinese and Russian. TermSuite's architecture is composed by 3-step modules: the Spotter, the Indexer and the Aligner. The Spotter module is responsible for preprocessing the input monolingual corpus, i.e., it performs tokenization, part-of-speech tagging, stemming and lemmatization. Then, the Indexer module uses both a statistic and a linguistic-based

approach to extract monolingual terminology from a monolingual corpus processed by the Spotter. Finally, the Aligner computes the translation of a source terminology into a target language. The source and target terms required are these already computed by the Indexer module, which means that the previous two steps should be repeated for the target language. The user can choose from several alignment options, such as the selection of the maximum number of translation candidates for a given source term, the use of similarity measures to compare the contexts of the term in the source and the target languages, amongst other advanced settings. Once all the parameters are set, it is possible to view and explore all the translation candidates ranked according to their similarity score within the tool or use the output XML file for other purposes.

3 Web-Based Terminology Extraction Tools

Although standalone TET still are predominant on todays TE applications market, the future web-based TE technologies will certainly evolve by migrating all standalone features to a web-based environment, which will allow them to consequently take over the leadership in the near future. As we will see, there are already some examples of this trend. The advantages are that web-based TET, compared to standalone tools, do not require any prior installation as they can be accessed within a web browser and that they make use of web technologies. Although most of web-based TET are often integrated as features in cutting-edge web-based applications with a wider purpose, such as managing corpora or terminology (e.g., Sketch Engine and Terminus, respectively), there also exist tools like the TET by Translated, which were developed with the proper purpose of terminology extraction.

Sketch Engine is an online tool created by Lexical Computing Ltd for building and managing corpora, which along with a number of corpus-processing features includes terminology extraction. It can be accessed under a paid commercial or academic license and supports 82 languages. This tool offers both monolingual and multilingual extraction. When extracting monolingual terminology, the user can choose whether to extract only single words (keywords) or multi-word terminological units (terms). In the output, the user can see the keywords or terms, links to the five most relevant Wikipedia articles for each of them, the term's score, its frequency in the searched corpus, and its frequency in the reference corpus. There are a variety of search options that can be tuned. For instance, the user can choose a different reference corpus, decide whether search for words or lemmas, and accentuate low or high-frequency keywords according to the preferences. The output can be downloaded as a TBX or CSV file. In order to perform multilingual term extraction the user needs to upload a TMX file with a parallel corpus aligned on the sentence or paragraph level. The terminology is first extracted within each language resulting in lists of candidate terms. In the second step, the system searches for such pairs of candidates which co-locate in the parallel documents most often. The resulting list of candidate pairs (terms in two languages) is then presented to the user. Results can be saved in a TBX or TXT file, which is especially convenient for computer-assisted translation tool users.

Translated s.r.l. a leading LSP developed a web-based tool that can be accessed directly on the company's website. It was created in order to help translators with their translation jobs by identifying the difficulties in the text and simplifying the process of creating glossaries. Up to the current date it supports only English, Italian and French. The system output includes the top 20 terms ranked by their score. In addition, the terms are given as hyperlinks to the corresponding Google search results. Below the list of terms the tool also shows all the terms in their full-sentence context. In order to easily differentiate the terms, each term is highlighted by a different color. In general, this tool is quite simple compared to the others, but can provide a fast and free solution any time it is needed.

Terminus is a web-based application for corpus and terminology management developed at the University Pompeu Fabra, Spain and it can be accessed by software licensing. The purpose of this tool is to integrate the complete process of terminographic work: textual corpus search, compilation and analysis, term extraction, glossary and project management, database

creation and maintenance, and dictionary edition. This is done with the help of a number of articulated modules, including the Analysis module, which has a semi-automatic term extraction feature. The extraction process has two options: the user can train a term extractor in a specific domain by incorporating an electronic dictionary containing terms of the same field, or simply apply a generic ready-to-use term extractor to any textual corpus. In addition, one can use other features to extract term candidates, such as the n-gram extractor, bi-gram extraction with association measures, keywords, and later manually validate relevant terms.

4 Frameworks

Frameworks are different from the other two types of tools because they are not complete software products but reusable software environments or libraries that can be used or even completely integrated in larger translation software applications, products or solutions. In particular, systems of this type are often used in information retrieval, where identification and indexing of terminology serves as an aid to information retrieval queries. In detail, the purpose of terminology extraction for both information retrieval and document retrieval is to isolate terms that contain enough informational content to support retrieval based on the queries supplied when querying a set of documents.

Keyphrase Extraction Algorithm (Kea) is a framework specially designed for automatically assigning terms to a document (aka keyphrase indexing). Kea is a platform-independent toolkit implemented in Java and distributed under the GNU General Public License. In detail, this framework can either be used for free indexing or for indexing with a controlled vocabulary. When used as free indexing, Kea looks for significant terms in a document. If on one hand, the free indexing option can be applied to any document and working language (as long as the corresponding stopword file and stemmer are provided). The controlled indexing, on the other hand, has the advantage that all documents are indexed in a consistent way disregarding their wording as the algorithm only collects those n-grams that match thesaurus terms.

Rainbow is a simple, yet powerful open-source platform-independent terminology extraction tool written in Java that uses statistic-based methods to automatically extract terms from multiple files and formats in any language. It is based on the Okapi Framework, a free, open-source and cross-platform framework that has a set of components and applications designed to help engineers, developers, translators and project managers involved in localization and translation-related tasks.

Java Automatic Term Extraction (JATE) is a JAVA toolkit that comprises several state-of-the-art term extractions algorithms. The motivation of this TET is three-fold: make available several automatic term extraction algorithms for the research community; encourage developers to built their methods under a uniform framework; and, enable comparative studies between different term extraction algorithms. JATE's workflow follows the typical TET steps: extract candidate terms from a corpus using linguistic tools; extract the candidates statistical features from the corpus; and, apply automatic terminology extraction algorithms to score the candidate terms domain representativeness based on their statistical features. So far, JATE's current version includes twelve state-of-the-art statistical algorithms.

5 Translators' preferences and opinions on the features of TET

As we mentioned above, translation is one of the most important applications of terminology extraction. However, it has not yet become a common part of the professional translation workflow. This was demonstrated by a user survey replied by over 600 translation professionals Zaretskaya et al. (2015), which showed that only 25% of the respondents regularly resorted to TE in their work. It could be due to unsatisfying performance of the existing tools, their interface design, or simply to translators' lack of awareness of these tools and of the benefits they can yield.

We have already seen that TET can differ as to various characteristics, such as their interface type (standalone, web-based or reusable libraries), document formats they support, languages they work with, as well as different search options.

	SDL Multiterm	*Simple Extractor*	*TermSuit*	*Sketch Engine*	*Translated s.r.l.*	*Terminus*	*Kea*	*Rainbow*	*JATE*
Bilingual extraction	✓		✓	✓					
Source and target context comparison	✓								
Terms validation	✓	✓		✓		✓	✓	✓	✓
Bilingual dictionaries compilation	✓		✓						
Context extraction	✓	✓	✓		✓	✓	✓	✓	✓
Support various file formats	✓	✓	✓	✓		✓	✓	✓	✓
Rank terms by frequency	✓	✓	✓	✓		✓	✓	✓	✓
Support for many languages	✓		✓	✓		✓	✓	✓	✓
Specify the minimal number of occurrences	✓	✓		✓		✓	✓	✓	✓
Show linguistic information	✓		✓			✓			
Specify the maximum number of translations			✓						
Stopword list option	✓	✓			✓		✓	✓	✓
Choose the minimum and maximum number of words per term	✓	✓					✓	✓	✓
Term statistics	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison chart of features for the selected tools.

According to the survey findings, 27% of the respondents preferred to have a TE feature within their computer-assisted translation (CAT) tool instead of a separate TE software. Some translators, however, preferred a web-based application (9%) or installing a standalone tool on their computer (8%). Nevertheless, the majority (56%) reported that they did not have any preference regarding the tool's interface. The fact that translators prefer to have a TE system integrated in their CAT tool is related to the general tendency of CAT tools to include more and more different features. Indeed, translators have to deal with a great number of tools that help them automatize different stages of the translation process, so they prefer having one tool with multiple functions rather than having to look for and in many cases pay for several tools.

Regarding the importance of the TE's features, the most useful feature according to survey's participants was bilingual term extraction. In fact, considering that within a translation workflow, terminology extraction is performed with the final objective to translate the extracted terms, it is more convenient to have the terms extracted in the two languages simultaneously. Bilingual extraction is much harder to perform than only monolingual as it requires a good word alignment system, so not many existing tools offer this feature. In particular, among the tools we considered in the previous section only SDL Multiterm Extract and Sketch Engine have bilingual extraction. Similarly, TermSuite also offers translation candidates for the extracted monolingual terms, which is a different procedure, but still leads to the same results: terms in two languages. The second ranked feature was the possibility to compare the context of the term in the source and the target language, which is another type of bilingual analysis suitable for the translation task. This feature is also quite rare, and of all the considered tools, only SDL Multiterm Extract allows such analysis. The possibility to validate terms or, in other words, choose the terms that should be extracted instead of extracting all terms was ranked third and is also considered useful for translators. This feature is offered by almost all systems, except for TermSuite and Translated. Compiling a bilingual dictionary from parallel texts is another useful feature, which is offered only by SDL Multiterm Extract and by TermSuite. Finally, the respondents considered it useful to extract context together

with terms or to see examples from the corpus. This is a common feature for many of the studied tools, including SDL Multiterm Extract, Simple Extractor or Translated.

Other features that were considered included support for different file formats, possibility to sort terms by frequency, support for many languages, possibility to specify the minimal number of occurrences of the words, show linguistic information about the term, and select the maximum number of translations for one term. All of them were considered useful, but were not among the most useful features.

And finally, some features were not considered so important by the respondents. One of them was the stopword list option: some of the tools, like Simple Extractor, allow to choose whether to use a stopword list, and others use it by default. Choosing the minimum and the maximum number of words per term, which was also among the least useful features, can be tuned by all the mentioned TE frameworks, for example. And finally, term statistics, which to some extent are provided by all tools, were not very important for most translators either. Table 1 shows which of the aforementioned features are presented in the 9 selected tools.

	Availability	Notes
SDL Multiterm	$500	Free demo available
Simple Extractor	$140	60-days demo
TermSuit	Open Source	
Sketch Engine	$65/ year	30-days demo
Translated s.r.l	Free	
Terminus	$440/ year	15-days demo
Kea	Open Source	
Raibow	Open Source	
JATE	Open Source	

Table 2: Depending on the purpose the quotes may vary. This table only shows the prices for licenses paid by individuals.

6 Conclusion

Although terminology extraction plays an important role in several disciplines such as linguistic research or language teaching, it is in the field of translation, particularly in the translation industry where its advantages are fully exploited and integrated in the workflow. An example of that is the use of bilingual term extraction, compiling dictionaries and comparing context in different languages as essential features for translators' work. In addition, it is also very useful for translators to see the terms in their context in order to understand their meaning and be able to find an adequate translation equivalent. Not all existing tools, however, provide these functionalities. We suggest that developing TET more suitable for the purpose of translation could help professionals in the industry take better advantage of TE technology. This has to be done, first of all, by taking into account the user requirements. As a step further in this direction, it would be necessary to investigate in more detail translators' attitudes towards TE tools. Especially, the reasons that prevent the vast majority of professional translators to adopt them. For instance, many translators might not be aware of their existence or understand their purpose, do not have time to learn how to use another complicated interface, or simply have other established procedures for dealing with terminology.

Acknowledgements

Hernani Costa and Anna Zaretskaya contributed equally to this work and are both supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n.317471.

References

Zaretskaya, A., Corpas Pastor, G., and Seghiri, M. (2015). Translators' requirements for translation technologies: a user survey. In *New Horizons in Translation and Interpreting Studies (Full papers)*, pages 247–254, Genebra, Switzerland. Tradulex.

Costa et al. (2016a)

Costa, H., Durán Muñoz, I., Corpas Pastor, G., and Mitkov, R. (2016b). **Compilação de Corpos Comparáveis Especializados: Devemos sempre confiar nas Ferramentas de Compilação Semi-automáticas?** *Linguamática*, 7(2):17.

Compilação de Corpos Comparáveis Especializados: Devemos sempre confiar nas Ferramentas de Compilação Semi-automáticas?

Compiling Specialised Comparable Corpora. Should we always thrust (Semi-)automatic Compilation Tools?

Hernani Costa
Universidade de Málaga
hercos@uma.es

Isabel Dúran Muñoz
Universidade de Málaga
iduran@uma.es

Gloria Corpas Pastor
Universidade de Málaga
g.corpas@uma.es

Ruslan Mitkov
Universidade de Wolverhampton
r.mitkov@wlv.ac.uk

Resumo

Decisões tomadas anteriormente à compilação de um corpo comparável têm um grande impacto na forma em que este será posteriormente construído e analisado. Diversas variáveis e critérios externos são normalmente seguidos na construção de um corpo, mas pouco se tem investigado sobre a sua distribuição de similaridade textual interna ou nas suas vantagens qualitativas para a investigação. Numa tentativa de preencher esta lacuna, este artigo tem como objetivo apresentar uma metodologia simples, contudo eficiente, capaz de medir o grau de similaridade interno de um corpo. Para isso, a metodologia proposta usa diversas técnicas de processamento de linguagem natural e vários métodos estatísticos, numa tentativa bem sucedida de avaliar o grau de similaridade entre documentos. Os nossos resultados demonstram que a utilização de uma lista de entidades comuns e um conjunto de medidas de similaridade distribucional são suficientes, não só para descrever e avaliar o grau de similaridade entre os documentos num corpo comparável, mas também para os classificar de acordo com seu grau de semelhança e, consequentemente, melhorar a qualidade do corpos através da eliminação de documentos irrelevantes.

Palavras chave

corpos comparáveis, linguística computacional, medidas de similaridade distribucional, compilação manual e semi-automática, processamento de linguagem natural.

Abstract

Decisions at the outset of compiling a comparable corpus are of crucial importance for how the corpus is to be built and analysed later on. Several variables and external criteria are usually followed when building a corpus but little is been said about textual distributional similarity in this context and the quality that it brings to research. In an attempt to fulfil this gap, this paper aims at presenting a simple but efficient methodology capable of measuring a corpus internal degree of relatedness. To do so, this methodology takes advantage of both available natural language processing technology and statistical methods in a successful attempt to access the relatedness degree between documents. Our findings prove that using a list of common entities and a set of distributional similarity measures is enough not only to describe and assess the degree of relatedness between the documents in a comparable corpus, but also to rank them according to their degree of relatedness within the corpus.

Keywords

comparable corpora, computational linguistics, distributional similarity measures, manual and semi-automatic compilation, natural language processing.

1 Introdução

O EAGLES - Expert Advisory Group on Language Engineering Standards Guidelines (EAGLES, 1996) define "corpos comparáveis" da seguinte forma: "Um corpo comparável é aquele que seleciona textos semelhantes em mais

de um idioma ou variedade. Devido à escassez de exemplos de corpos comparáveis, ainda não existe um acordo sobre a sua similaridade.". Desde o momento em que esta definição foi criada em 1996, muitos corpos comparáveis foram compilados, analisados e utilizados em várias disciplinas. A verdade é que este recurso acabou por se tornar essencial em várias áreas de investigação, tais como o Processamento de Linguagem Natural (PLN), terminologia, ensino de idiomas e tradução automática e assistida, entre outras. Neste momento podemos afirmar que não existe mais "escassez de exemplos de corpos comparáveis". Como Maia (2003) referiu: "os corpos comparáveis são vistos como uma resposta às necessidades de textos como exemplo de texto 'natural' original na cultura e idioma de origem" e, portanto, não é surpresa nenhuma que tenhamos assistido a um aumento no interesse por esses recursos e, um grande impulso na compilação de corpos comparáveis, especialmente no campo da investigação nas últimas décadas.

Contudo, de momento, "ainda não existe um acordo sobre a sua similaridade". A incerteza sobre os dados com que estamos a lidar ainda é um problema inerente para aqueles que lidam com corpos comparáveis. De facto, pouca investigação tem sido feita sobre a caraterização automática deste tipo de recurso linguístico, e tentar fazer uma descrição significativa do seu conteúdo é, muitas vezes, uma tarefa no mínimo arriscada (Corpas Pastor e Seghiri, 2009). Geralmente a um corpo é atribuído uma breve descrição do seu conteúdo, como por exemplo "transcrições de falas casuais" ou "corpo especializado comparável de turismo", juntamente com outras etiquetas que descrevem a sua autoria, data de criação, origem, número de documentos, número de palavras, etc. Na nossa opinião, estas especificações são de pouca valia para aqueles que procuram um corpo representativo de um domínio específico de elevada qualidade, ou até mesmo para aqueles que pretendem reutilizar um determinado corpo para outros fins. Desta forma, a maioria dos recursos à nossa disposição são construídos e partilhados sem que seja feita uma análise profunda ao seu conteúdo. Aqueles que os utilizam cegamente, confiam nas pessoas ou no grupo de investigação por detrás do seu processo de compilação, sem que conheçam a verdadeira qualidade interna do recurso, ou por outras palavras, sem conhecimento real sobre a quantidade de informação partilhada entre os seus documentos, ou quão semelhantes os documentos são entre si.

Assim, este trabalho tenta colmatar esta lacuna propondo uma nova metodologia que poderá ser utilizada em corpos comparáveis. Depois de selecionar o corpo que irá ser usado como cobaia em várias experiências, apresentamos a metodologia que explora várias técnicas de PLN juntamente com várias Medidas de Similaridade Distribucional (MSD). Para este efeito usámos uma lista de entidades comuns como parâmetro de entrada das MSD. Assumindo que os valores de saída das várias MSD podem ser usados como unidade de medida para identificar a quantidade de informação partilhada entre os documentos, a nossa hipótese é que estes valores possam ser posteriormente utilizados para descrever e caracterizar o corpo em questão.

O resto do artigo está estruturado da seguinte forma. A secção 2 descreve as vantagens e as desvantagens da compilação manual e automática de corpos e revela as atuais tendências de investigação usadas na compilação automática de corpos comparáveis. A secção 3 introduz alguns conceitos fundamentais relacionados com as MSD, ou seja, explica os fundamentos teóricos, trabalhos relacionados e as medidas utilizadas neste trabalho. A secção 4 apresenta o corpo utilizado nas nossas experiências, enquanto que a secção 5 descreve em detalhe a metodologia proposta, juntamente com todas as ferramentas, bibliotecas e *frameworks* utilizadas. E, finalmente, antes das conclusões finais (secção 7), a secção 6 descreve em detalhe os resultados obtidos.

2 Compilação Manual vs. Compilação Semi-automática

A compilação automática ou semi-automática de corpos comparáveis (ou seja, corpos compostos por textos originais semelhantes num ou mais idiomas usando os mesmos critérios de *design* (EAGLES, 1996; Corpas Pastor, 2001)) têm demonstrado muitas vantagens para a investigação atual, reduzindo particularmente o tempo necessário para construir um corpo e aumentando a quantidade de textos recuperados. Com ferramentas automáticas de compilação como o BootCaT (Baroni e Bernardini, 2004), WebBootCaT (Baroni et al., 2006) ou o Babouk (de Groc, 2011), hoje em dia é possível construir um corpo de grande tamanho num reduzido período de tempo, em contraste com o demorado protocolo de compilação e o número limitado de textos recuperados no mesmo intervalo de tempo quando a compilação é realizada manualmente.

De facto, publicações recentes demonstram que a compilação automática está a superar a compilação manual, sendo cada vez maior o número de investigadores que tiram partido de ferramentas de compilação automática na construção dos seus corpos (Barbaresi, 2014; Jakubíček et al., 2014; Barbaresi, 2015; H. El-Khalili, Haddad e El-Ghalayini, 2015). A verdade é que neste momento é um truísmo dizer que a compilação automática de corpos está a ganhar terreno sobre a compilação manual.

Apesar de ser possível compilar mais rapidamente maiores corpos comparáveis num curto espaço de tempo – o que é sem dúvida a maior vantagem da compilação automática – é contudo necessário analisar todo o espectro de propriedades implícitas no processo. Em primeiro lugar, um dos inconvenientes mais importantes a considerar quando se lida com a compilação automática é o ruído, ou seja, a quantidade de informação irrelevante que acaba por ser adicionada ao corpo durante o processo. Ruído este que se tenta colmatar através de uma supervisão rigorosa nas primeiras fases, de modo a evitar possíveis repercussões nas fases seguintes. Deste modo, é quase desnecessário afirmar que a compilação automática também requer intervenção humana a fim de obter bons resultados durante o processo de compilação - daí a origem da palavra "semi-automática". Contudo, esta intervenção torna-se uma tarefa bastante tediosa e cansativa, dada a necessidade de filtrar determinados domínios na rede, eliminar pares de entidades ou páginas na rede irrelevantes oferecidas pela ferramenta de compilação (Gutiérrez Florido, Corpas Pastor e Seghiri, 2013).

Outra caraterística interessante de analisar é o grau de semelhança entre documentos compilados manualmente e semi-automaticamente. Apesar de à primeira vista pensarmos que a compilação manual é a única que garante a qualidade em termos de forma e conteúdo num corpo, devido ao facto deste tipo de compilação ser mais minuciosa em termos de seleção dos textos a serem adicionados ao corpo, até ao momento ainda não existe um método formal que prove a sua veracidade. Sendo a forma e conteúdo de suma importância na construção de corpos comparáveis, e posteriormente na análise do mesmo, este trabalho tem como principal objetivo propor um método capaz de descrever, medir e classificar em termos de forma e conteúdo o grau de similaridade em corpos comparáveis. Noutras palavras, capaz de avaliar o grau de semelhança/ similaridade dentro de um corpo compilado manualmente ou semi-automaticamente. E assim permitir que o investigador responsável pela compilação tenha um conhecimento mais aprofundado sobre os documentos com que está a lidar para que possa posteriormente decidir quais devem ou não fazer parte do corpo.

Numa tentativa de estandardizar o nosso trabalho, e considerando as limitações de cada tipo de compilação, tivemos em conta vários fatores comuns que devem ser satisfeitos por ambos tipos de compilação. Estas variáveis devem ser estabelecidas de modo a garantir a fiabilidade do corpo, a sua coerência interna e a representatividade do domínio. Deste modo, Bowker e Pearson (2002) propõe vários critérios a serem seguidos, os quais estão relacionados com as línguas de trabalho e o nível de especialização. Em seguida enumeramos os vários critérios externos a serem considerados:

- Critério temporal: a data de publicação ou criação dos textos selecionados;
- Critério geográfico: origem geográfica dos textos;
- Critério formal: autenticidade dos textos completos ou fragmentados;
- Tipologia dos textos: o género textual a que os textos pertencem;
- Critério de autoria: a fonte dos textos (autor, instituição, etc.).

É importante referir que, de modo a garantir a homogeneidade do corpo usado neste trabalho, estes critérios foram seguidos durante o processo de compilação, como explicado na secção 4. Além disso, é também importante referir que neste trabalho ambas as abordagens (manual ou semi-automática) usam as mesmas ferramentas para recuperar documentos (ou seja, o mesmo motor de busca).

3 Medidas de Similaridade Distribucional (MSD)

Embora a tarefa de estruturar informação a partir de linguagem natural não estruturada não seja uma tarefa fácil, o Processamento de Linguagem Natural (PLN) em geral e, Recuperação de Informação (RI) (Singhal, 2001) e Extração de Informação (EI) (Grishman, 1997) em particular, têm melhorado o modo como a informação é acedida, extraída e representada. Em particular, RI e EI desempenham um papel crucial na tarefa de localizar e extrair informação específica de uma coleção de documentos ou outro tipo de recursos em linguagem natural, de

acordo com um determinado critério de busca. Para isso, estas duas áreas do conhecimento tiram partido de vários métodos estatísticos para extrair informação sobre as palavras e suas coocorrências. Essencialmente, esses métodos visam encontrar as palavras mais frequentes num documento e usar essa informação como atributo quantitativo num determinado método estatístico. Partindo do teorema distribucional de Harris (1970), o qual assume que palavras semelhantes tendem a ocorrer em contextos semelhantes, esses métodos estatísticos são adequados, por exemplo, para encontrar frases semelhantes com base nas palavras contidas nas mesmas (Costa et al., 2015), ou, por exemplo, para extrair e validar automaticamente entidades semânticas extraídas de corpos (Costa, Gonçalo Oliveira e Gomes, 2010; Costa, 2010; Costa, Gonçalo Oliveira e Gomes, 2011). Para este efeito, assume-se que a quantidade de informação contida, por exemplo, num determinado documento poderá ser acedida através da soma da quantidade de informação contida nas palavras do mesmo. Além disso, a quantidade de informação transmitida por uma palavra pode ser representada pelo peso que lhe é atribuído (Salton e Buckley, 1988). Deste modo, o Spearman's Rank Correlation Coefficient (SCC) e o Chi-Square (χ^2), duas medidas frequentemente aplicadas na área de RI, podem ser utilizadas para calcular a similaridade entre dois documentos escritos no mesmo idioma (ver secção 3.1 e 3.2 para mais detalhes sobre estas medidas). Ambas as medidas são particularmente úteis para este trabalho, visto que ambas são: independentes do tamanho do texto (ambas usam uma lista das entidades comuns); e, independentes do idioma.

Devido a ser independente do tamanho dos textos e à sua simplicidade de implementação, a medida distribucional do SCC tem demonstrado a sua eficácia no cálculo da similaridade entre frases, documentos e até mesmo em corpos de tamanhos variados (Costa et al., 2015; Costa, 2015; Kilgarriff, 2001).

A medida de similaridade do χ^2 também tem demonstrado a sua robustez e alto desempenho. A título de exemplo, o χ^2 tem vindo a ser utilizado para analisar o componente de conversação no Corpo Nacional Britânico (Rayson, Leech e Hodges, 1997), para comparar corpos (Kilgarriff, 2001), e até mesmo para identificar grupos de tópicos relacionados em documentos transcritos (Ibrahimov, Sethi e Dimitrova, 2002). Embora seja uma medida estatística simples, o χ^2 permite avaliar se a relação entre duas variáveis numa amostra é devida ao acaso, ou, pelo contrário, a relação é sistemática.

Devido às razões mencionadas anteriormente, as Medidas de Similaridade Distribucional (MSD), em geral, e o SCC e χ^2 em particular, têm uma vasta gama de aplicabilidades (Kilgarriff, 2001; Costa, 2015; Costa, Corpas Pastor e Mitkov, 2015). Deste modo, este trabalho tem como objetivo provar que estas medidas simples, contudo robustas e de alto desempenho, permitem descrever o grau de similaridade entre documentos em corpos especializados. Em seguida descrevemos em detalhe como funcionam estas duas MSD.

3.1 Spearman's Rank Correlation Coefficient (SCC)

Neste trabalho o Spearman's Rank Correlation Coefficient (SCC) é utilizado e calculado do mesmo modo que no artigo do Kilgarriff (2001). Inicialmente é criada uma lista de entidades comuns[1] L entre dois documentos d_l e d_m, onde $L_{d_l,d_m} \subseteq (d_l \cap d_m)$. É possível usar n entidades comuns ou todas as entidades comuns entre dois documentos, onde n corresponde ao total número de entidades comuns em $\mid L \mid$, ou seja, $\{n \mid n \in \mathbb{N}^0, n \leq \mid L \mid\}$ – neste trabalho são utilizadas todas as entidades comuns encontradas entre dois documentos, ou seja, $n = \mid L \mid$. Em seguida, por cada documento, as listas de entidades comuns (por exemplo, L_{d_l} and L_{d_m}) são ordenadas por ordem crescente de frequência ($R_{L_{d_l}}$ e $R_{L_{d_m}}$), ou seja, a entidade menos frequente recebe a posição 1 no ranking e a entidade mais frequente recebe a posição n. Em caso de empate, onde mais do que uma entidade aparece no documento o mesmo número de vezes, é atribuída a média das posições. Por exemplo, se as entidades e_a, e_b e e_c ocorrerem o mesmo número de vezes e as suas posições forem 6, 7 e 8, a todas elas é atribuída a mesma posição no ranking, ou seja, a sua nova posição no ranking seria $\frac{6+7+8}{3} = 7$. Finalmente, para cada entidade comum $\{e_1, ..., e_n\} \in L$ em cada um dos documentos é calculada a diferença entre as suas posições e posteriormente normalizada através da soma dos quadros das suas diferenças $\left(\sum_{i=1}^{n} s_i^2\right)$. A equação completa do SCC é apresentada na Expressão 1, onde $\{SCC \mid SCC \in \mathbb{R}, -1 \leq SCC \leq 1\}$.

Como exemplo, imagine-se que e_x é

[1] Neste trabalho, o termo "entidade" refere-se a "palavras simples", as quais podem ser um *token*, um lema ou um stem.

uma entidade comum (ou seja, $\{e_x\} \in L$) e, $R_{L_{d_l}} = \{1\#e_{n_{d_l}}, 2\#e_{n-1_{d_l}}, ..., n\#e_{1_{d_l}}\}$ e $R_{L_{d_m}} = \{1\#e_{n_{d_m}}, 2\#e_{n-1_{d_m}}, ..., n\#e_{1_{d_m}}\}$ são as listas ordenadas de entidades comuns de d_l e d_m, respetivamente. Assumindo que e_x é o $3\#e_{n-2_{d_l}}$ e $1\#e_{n_{d_m}}$, ou seja, e_x está na posição 3 do ranking em $R_{L_{d_l}}$ e na posição 1 em $R_{L_{d_m}}$, s seria calculado da seguinte forma: $s^2_{e_x} = (3-1)^2$ e, o resultado seria 4. Em seguida este processo seria repetido para as restantes $n-1$ entidades e o resultado do SCC corresponderia ao valor de similaridade entre d_l e d_m.

$$SCC(d_i, d_j) = 1 - \frac{6 * \sum_{i=1}^{n} s_i^2}{n^3 - n} \quad (1)$$

3.2 Chi-Square (χ^2)

A medida do Chi-square (χ^2) também usa uma lista de entidades comuns (L). E à semelhança do SCC, também é possível usar n entidades comuns ou todas as entidades comuns entre dois documentos. Também neste caso optámos por usar a lista completa, ou seja, todas as entidades comuns encontradas entre dois documentos ($n =\mid L \mid$). O número de ocorrências de uma determinada entidade em L, que seria expectável em cada um dos documentos, é calculado usando a lista de frequências. Se o tamanho do documento d_l e d_m forem N_l e N_m e a entidade e_i tiver as seguintes frequências observadas $O(e_i, d_l)$ e $O(e_i, d_m)$, então os valores esperados seriam $e_{i_{d_l}} = \frac{N_l * (O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$ e $e_{i_{d_m}} = \frac{N_m * (O(e_i,d_l)+O(e_i,d_m))}{N_l+N_m}$. Na equação 2 é apresentada a fórmula completa do χ^2, onde O corresponde ao valor da frequência observada e E a frequência esperada. Assim, o valor resultante do χ^2 deverá ser interpretado como a distância interna entre dois documentos. Também é importante referir que $\{\chi^2 \mid \chi^2 \in \mathbb{R}, 1 \leq \chi^2 < +\infty\}$, o que significa que quanto menos relacionadas as entidades forem em L, menor será o valor do χ^2.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad (2)$$

A Tabela 1 apresenta um exemplo de uma tabela de contingências. Assumindo que existem duas entidades comuns e_i e e_j entre dois documentos d_l e d_m (ou seja, $L = \{e_i, e_j\}$), esta tabela apresenta: i) as frequências observadas (O); ii) os totais nas margens; iii) as frequências esperadas (E), que foram obtidas através da seguinte fórmula: $\frac{column\ total}{N} * row\_total$, por exemplo, $E(e_i, d_l) = \frac{14}{26} * 15 = 8.08$. Assim que calculadas as frequências esperadas, o próximo passo seria calcular o χ^2 (veja-se Equação 3).

	d_l	d_m	*Total*
e_i	O=11 E=8.08	O=4 E=6.92	15
e_j	O=3 E=5.92	O=8 E=5.08	11
Total	14	12	26

Tabela 1: Exemplo de uma tabela de contingência.

$$\frac{(11-8.08)^2}{8.08} + \frac{(3-5.92)^2}{5.92} + \frac{(4-6.92)^2}{6.92} + \frac{(8-5.08)^2}{5.08} = 5.41 \quad (3)$$

4 O Corpo INTELITERM

O corpo INTELITERM[2] é um corpo comparável especializado composto por documentos recuperados da Internet. Inicialmente foi compilado manualmente, por investigadores, com o objetivo de construir um corpo em inglês, espanhol, alemão e italiano livre de ruído e representativo na área do Turismo e Beleza. No entanto, numa fase posterior, a fim de aumentar o tamanho do mesmo, mais documentos foram recuperados automaticamente usando a ferramenta de compilação BootCaT[3] (Baroni e Bernardini, 2004). De modo a manter a homogeneidade e a qualidade do corpo, em ambos os processos de compilação foram seguidas as mesmas variáveis e critérios externos (ver Tabela 2).

Em detalhe, o corpo comparável INTELITERM pode ser dividido em quatro subcorpos de acordo com o idioma, ou seja, inglês, espanhol, alemão e italiano. Estes subcorpos, por sua vez podem ser subdivididos por tipo de documento, isto é, textos originais compilados manualmente, textos traduzidos compilados manualmente e textos originais compilados automaticamente. Dado o reduzido tamanho do corpo (veja-se Tabela 3), decidimos usar todos os seus documentos, ou seja, todos os documentos originais e traduzidos compilados manualmente para o inglês (*i_en_od* e *i_en_td*), espanhol (*i_es_od* e *i_es_td*), alemão (*i_de_od* e *i_de_td*) e italiano (*i_it_od* - os investigadores não

[2] http://www.lexytrad.es/proyectos.html

[3] `http://bootcat.sslmit.unibo.it`

Critério	Descrição
Temporal	A data de publicação ou criação dos textos selecionados deve ser tão recente quanto possível.
Geográfico	De modo a evitar uma possível variação terminológica diatópica, como o espanhol falado no México ou Venezuela, todos os textos selecionados são geograficamente limitados, ou seja, todos os textos utilizados, por exemplo, em espanhol são provenientes de Espanha, e todos os textos italianos são da Itália.
Formal	Os textos selecionados referem-se a um contexto de comunicação especializado, ou seja, a um contexto de nível médio-alto de especialização, são originalmente escritos nas línguas do estudo e estão no seu formato eletrónico original.
Género ou tipologia textual	Todos os textos selecionados pertencem ao mesmo género, ou seja, são textos promocionais recuperados da Internet contendo informação sobre produtos e serviços de bem-estar e beleza na área do turismo.
Autor	Todos os textos são documentos autênticos criados por autores relevantes, instituições ou empresas.

Tabela 2: Variáveis e critérios externos utilizados durante o processo de compilação.

encontraram textos traduzidos para o italiano), assim como todos os documentos compilados automaticamente usando a ferramenta de compilação automática *b*oot*c*aT para o inglês, espanhol, alemão e italiano (*bc_en*, *bc_es*, *bc_de* and *bc_it*, respetivamente). Toda a informação relativa aos subcorpos referidos anteriormente é apresentada na Tabela 3. Esta tabela apresenta o número de documentos (nD), o número de palavras únicas (*types*), o número total de palavras (*tokens*), a relação entre palavras únicas e o número total de palavras ($\frac{types}{tokens}$) por subcorpos e o tipo de fonte (sT), a qual pode ser original, tradução ou *crawled*/ recuperado automaticamente (ori., trans. e craw., respetivamente). Os valores apresentados na Tabela 3 foram obtidos através da ferramenta de análise de concordância Antconc 3.4.3 (Anthony, 2014).

	nD	types	tokens	$\frac{types}{tokens}$	sT
i_en_od	151	11.6k	496.2k	0,023	ori.
i_en_td	60	6.9k	83.1k	0,083	trans.
i_es_od	224	13.0k	207.3k	0,063	ori.
i_es_td	27	3.4k	16.4k	0,207	trans.
i_de_od	138	21.4k	199.8k	0,049	ori.
i_de_td	109	5.5k	26.8k	0,205	trans.
i_it_od	150	19.9k	386.2k	0,051	ori.
bc_en	111	41.1k	563.5k	0,073	craw.
bc_es	246	32.8k	735.4k	0,045	craw.
bc_de	253	58.3k	482.4k	0,121	craw.
bc_it	122	11.9k	81.5k	0,147	craw.

Tabela 3: Informação estatística dos vários subcorpos do INTELITERM.

5 Medindo o Grau de Similaridade entre Documentos

Esta secção tem como objetivo apresentar uma metodologia simples, contudo eficiente capaz de descrever e extrair informação sobre o grau interno de similaridade de um determinado corpo. De facto, em última instância, esta metodologia permitir-nos-á não só descrever os documentos num corpo, mas também medir e classificar documentos com base nos seus valores de similaridade. Em seguida descrevemos a metodologia usada para este fim, juntamente com todas as ferramentas, bibliotecas e *frameworks* utilizadas no processo.

i) **Pré-processamento dos dados**: em primeira instância processámos o corpo com o OpenNLP[4] de modo a delimitar as frases e as palavras. Relativamente ao processo de anotação, utilizámos o TT4J[5], uma biblioteca em Java que permite invocar a ferramenta TreeTagger (Schmid, 1995) – uma ferramenta criada especificamente para identificar a categoria gramatical e o lema das palavras. Em relação ao *stemming*, usámos o algoritmo Porter stemmer fornecido pela biblioteca Snowball[6]. Também foi implementado manualmente um módulo para remover sinais de pontuação e caracteres especiais dentro das palavras. Além disso, de modo a eliminarmos o ruído, foi criada uma lista de stopwords[7] para identificar as palavras mais frequentes no corpo, ou seja, palavras vazias sem informação semântica. Uma vez processado um determinado documento, ou seja, depois de delimitar as frases, identificar as palavras, a sua categoria gramatical, o seu lema e o seu stem, o sistema cria um novo ficheiro onde é guardada toda esta

[4] https://opennlp.apache.org
[5] http://reckart.github.io/tt4j/
[6] http://snowball.tartarus.org
[7] Disponíveis através do seguinte endereço na rede: https://github.com/hpcosta/stopwords.

nova informação. Além disso, também é adicionado ao ficheiro um vetor booleano que descreve se uma entidade é uma palavra irrelevante (ou seja, stopword) ou não. Desta forma, o sistema irá ser capaz de utilizar somente as palavras, lemas e stems que não sejam stopwords.

ii) **Identificação da lista de entidades comuns entre documentos**: de modo a identificar a lista de entidades comuns (para futura referência, EC), foi criada uma matriz de coocorrências por cada par de documentos. Neste trabalho, somente pares de documentos com pelo menos uma entidade em comum são processados. Como exigido pelas MSD (ver secção 3), a frequência das EC em ambos os documentos são guardadas numa matriz de coocorrências ($L_{d_l,d_m} = \{e_i, (f(e_i,d_l), f(e_i,d_m)); e_j, (f(e_j,d_l), f(e_j,d_m)); ...; e_n, (f(e_n,d_l), f(e_n,d_m))\}$, onde f representa a frequência de uma entidade num determinado documento d). Com o objetivo de analisar e comparar o desempenho das várias MSD foram criadas três listas para serem utilizadas como parâmetros de entrada: a primeira usando o número de tokens em comum (NTC), a segunda usando o número de lemas em comum (NLC) e a terceira usando o número de stems em comum (NSC).

iii) **Calcular a similaridade entre documentos**: a similaridade entre documentos foi calculada aplicando as várias MSD ($MSD = \{MSD_{EC}, MSD_{SCC}, MSD_{\chi^2}\}$, onde $_{EC}$, $_{SCC}$ e $_{\chi^2}$ correspondem ao número de entidades comuns ao Spearman's Rank Correlation Coefficient e ao Chi-Square, respetivamente), usando os três parâmetros de entrada (NTC, NLC e NSC).

iv) **Calcular a pontuação final do documento**: a pontuação final do documento $MSD(d_l)$ resulta da média das similaridades entre o documento d_l com todos os demais documentos na coleção de documentos, ou seja, $MSD(d_l) = \frac{\sum_{i=1}^{n-1} MSD_i(d_l,d_i)}{n-1}$, onde n representa o número total de documentos na coleção e $MSD_i(d_l,d_i)$ o valor de similaridade entre o documento d_l com o documento d_i.

v) **Classificar os documentos**: por fim, os documentos são classificados por ordem descendente de acordo com o valor resultante final das várias MSD (ou seja, MSD_{EC}, MSD_{SCC} ou MSD_{χ^2}).

6 Avaliando o Corpo usando MSD

Depois de apresentado o problema que pretendemos explorar, a metodologia que iremos aplicar e os dados com os quais iremos trabalhar, é hora de juntar todas as peças num cenário de teste e explicar as nossas descobertas. Para este efeito, as Medidas de Similaridade Distribucional (MSD), apresentados na secção 3, serão aplicadas para explorar e classificar os documentos do corpo INTELITERM. Esta experiência divide-se em duas partes distintas. Na primeira parte, usaremos os vários subcorpos compilados manualmente para explorar e comparar o conteúdo dos documentos originais com os traduzidos, de modo a compreender como eles diferem entre si de um ponto de vista estatístico (secção 6.1). Depois, na segunda parte, faremos uma análise comparativa entre os documentos compilados manualmente com os semi-automaticamente compilados (secção 6.2). Por fim, esta secção termina com uma discussão geral sobre os resultados obtidos (secção 6.3).

A fim de descrever os dados em mãos é aplicada a metodologia apresentada na secção 5, juntamente com as três diferentes MSD, ou seja: o número de entidades comuns (EC); o Spearman's Rank Correlation Coefficient (SCC); e o Chi-Square (χ^2). Como parâmetro de entrada para as diferentes MSD, usaremos três diferentes listas de entidades (isto é, tokens, lemas e stems). As Figuras 1, 2 e 3 apresentam o número médio (av) do número de tokens comuns (NTC) entre documentos, os valores resultantes do SCC e do χ^2, juntamente com os seus desvios padrão correspondentes (σ - linhas verticais que se estendem a partir das barras) por medida e subcorpos (ou seja, documentos *o*riginais, *t*raduzidos e compilados automaticamente com o *boot*caT. Usaremos os seus acrónimos, a partir deste momento: *i_od*, *i_td* and *bc*, respetivamente).

É importante referir que neste trabalho usamos todos os documentos do corpo INTELITERM e, portanto, todos os resultados observados resultam de toda a população, e não de uma amostra. Ou seja, são utilizados todos os documentos em: inglês (*i_en_od*, *i_en_td* e *bc_en*); espanhol (*i_es_od*, *i_es_td* e *bc_es*); alemão (*i_de_od*, *i_de_td* e *bc_de*); e italiano (*i_it_od* e *bc_it*) - importante referir novamente que para o italiano não existe um o subcorpo de

documentos traduzidos (ver secção 4).

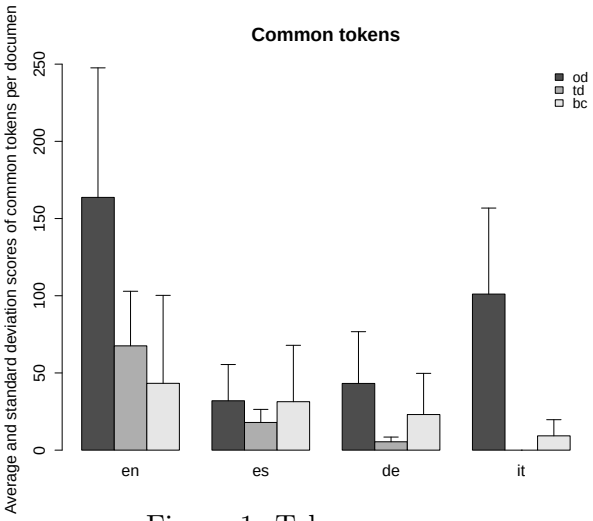


Figura 1: Tokens comuns.

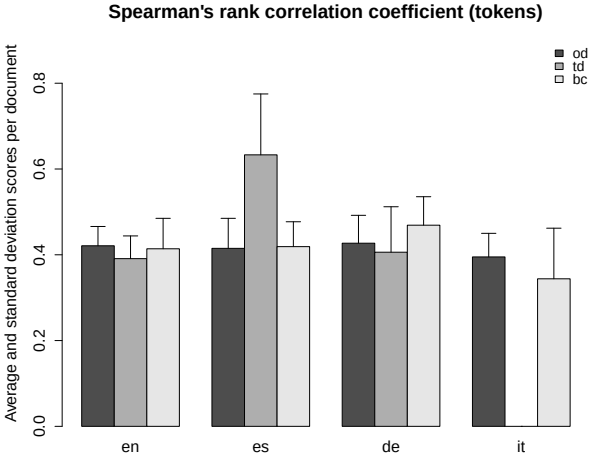


Figura 2: SCC.

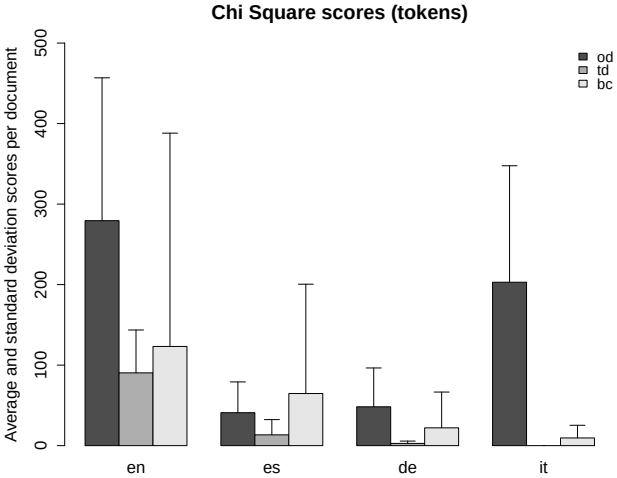


Figura 3: χ^2.

6.1 Documentos Originais vs. Traduzidos

As Figuras 4 a 12 apresentam os valores médios por documento num formato de *box plot* para todas as combinações MSD *vs.* subcorpo. Em cada uma das *box plot* é apresentada a gama de variação (mínimo e máximo), o intervalo de variação (variação interquartil), a mediana e os valores mínimos e máximos extremos (também conhecidos como *outliers*).

A primeira observação que podemos fazer a partir das Figuras 4, 7 e 10 é que as distribuições entre os distintos parâmetros de entrada são bastante semelhantes. Embora não seja possível generalizar estes resultados para outros tipos de corpos ou domínios, todas as MSD sugerem a mesma conclusão: é possível alcançar resultados aceitáveis apenas usando tokens, ou seja, palavras na sua forma original. Como os stems e os lemas exigem mais poder computacional e tempo para serem processados - especialmente os lemas, devido à sua dependência à categoria gramatical e ao tempo de processamento subjacente – a possibilidade de usar apenas tokens é uma mais valia não só para as MSD, mas principalmente para o método proposto neste trabalho.

Deste modo vamo-nos focar nas Figuras 4, 5 e 6. Com base nos resultados apresentados nas mesmas, podemos afirmar que os valores obtidos por cada subcorpo é simétrico (distribuição simétrica com a mediana no centro do retângulo), o que significa que os dados seguem uma distribuição normal. Contudo, há algumas exceções, como por exemplo nos valores médios para o SCC e para o χ^2, mais precisamente para o subcorpo *i_es_td* e para o *i_de_td*, os quais serão mais tarde analisados em detalhe nesta secção. Outra observação interessante está relacionada com o elevado número de entidades comuns (EC) - veja-se Figuras 1, 4, 7 e 10 - nos documentos originais (*i_en_od*, *i_es_od* e *i_de_od*) quando comparado com os documentos traduzidos (*i_en_td*, *i_es_td* e *i_de_td*, respetivamente). Por exemplo, o subcorpo *i_en_od* (o subcorpo em inglês que contém documentos originais) contém 163,70 tokens em comum por documento em média (av) com um desvio padrão (σ) de 83,89, enquanto que o subcorpo *i_en_td* (o qual contém textos traduzidos em inglês) tem somente 67,54 tokens comuns por documento em média com um σ=35,35 (ver Figura 1). A mesma observação pode ser feita para os subcorpos originais em espanhol e alemão (*i_es_od*={av=31,97; σ=23,48} e *i_de_od*={av=43,21; σ=33,52}) com os seus subcorpos traduzidos (*i_es_td*={av=17,93;

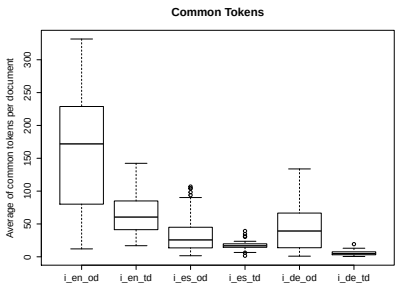


Figura 4: NTC.

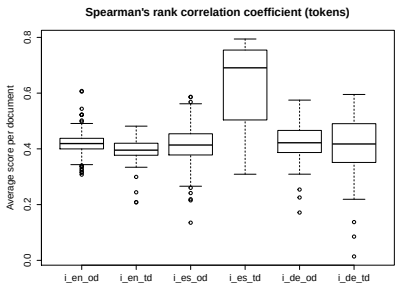


Figura 5: SCC.

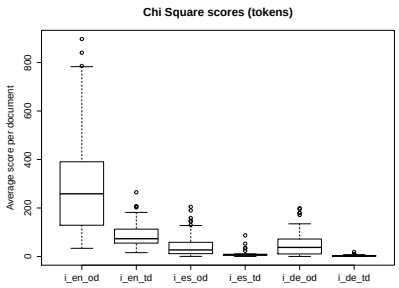


Figura 6: χ^2.

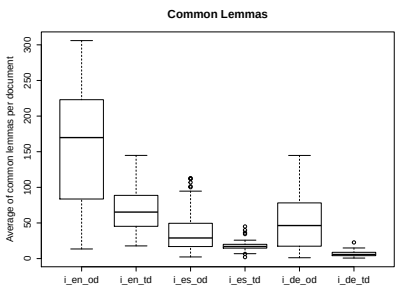


Figura 7: Lemas.

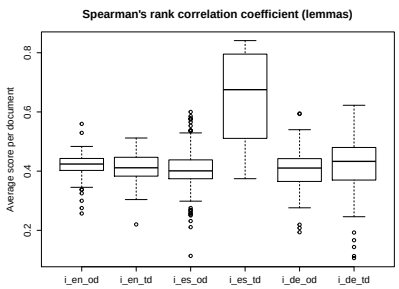


Figura 8: SCC.

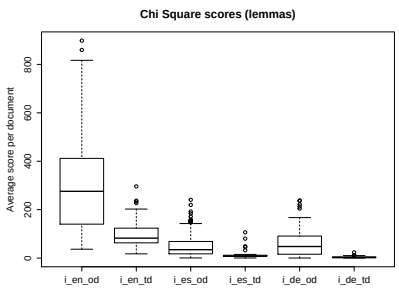


Figura 9: χ^2.

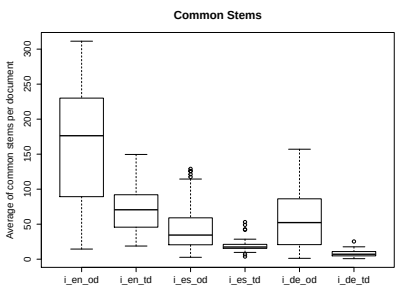


Figura 10: Stems.

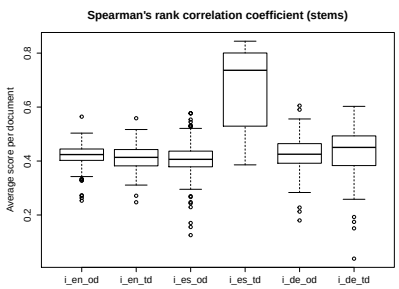


Figura 11: SCC.

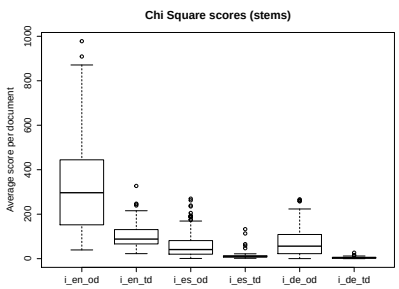


Figura 12: χ^2.

σ=8,46} e *i_de_td*={av=5,42; σ=3,05}), ver Figuras 1 e 4 - repare-se que a Figura 4 mostra como os dados estão distribuídos acima e abaixo da mediana e a Figura 1 apresenta as distintas médias e seus desvios padrão correspondentes. Uma possível explicação para estes valores baseia-se no fato destes documentos, recuperados da Internet, serem documentos traduzidos (ou seja, traduzidos de diferentes línguas e por diferentes tradutores) e, consequentemente, devido à variabilidade das várias caraterísticas linguísticas, tais como vocabulário, estilo, repetição, fontes, etc., em cada um dos documentos, pode muito bem explicar o porquê de haver um menor número de EC entre os documentos traduzidos quando comparado com os documentos originais.

Embora a média do número de tokens comuns por documento (NTC) seja maior para o corpo *i_en_od*, a amplitude inter-quartis (IQR) é maior que nos demais subcorpos (ver Figuras 1 e 4), o que significa que em média, 50% dos dados estão mais distribuídos e, consequentemente, a média de NTC por documento é mais variável. Além disso, na Figura 4 podemos verificar que os *whiskers* são longos (ou seja, as linhas que se estendem verticalmente a partir do retângulo), o que poderá indicar uma certa variabilidade fora dos quartis superiores e inferiores (ou seja entre o máximo e o Q3 e entre o Q1 e o mínimo). Portanto, podemos dizer que o subcorpo *i_en_od* contêm uma grande variedade de tipos de documentos e, consequentemente, alguns deles estão minimamente correlacionados com os demais documentos do subcorpo. No entanto, os dados são positivamente assimétricos, o que significa que a maioria está fortemente correlacionada, isto é, os documentos partilham um elevado NTC entre si. Esta ideia pode ser sustentada pelos valores médios do SCC e o elevado número de *outliers* positivos que se observam na Figura 5. Além disso, a média de 0,42 para o SCC e σ=0,045 também corroboram a existência de uma forte correlação entre os documentos no subcorpo *i_en_od*. Em relação aos

valores do χ^2, o longo *whisker* que sai do Q1, na Figura 6, também deve ser interpretado como indício de um elevado grau de similaridade entre os documentos.

Em relação ao subcorpo *i_en_td*, os valores do NTC, do SCC e do χ^2 (Figuras 4, 5 e 6) e, a média de 67,54 tokens comuns por documento e o σ=35,35 (Figura 1) sugerem que os dados estão normalmente distribuídos (Figura 5) e os documentos - não tanto como no subcorpo *i_en_od*, contudo - também estão fortemente relacionados entre si.

De todos os subcorpos, o *i_es_od* é o maior, contendo 224 documentos (Tabela 3). No entanto, as Figuras 1 e 4 revelam que o NTC é mais baixo em comparação com os dois subcorpos em inglês. Embora uma análise linguística mais aprofundada nos daria uma explicação mais precisa, uma possível teoria passa pelo facto de que o espanhol tem uma morfologia mais rica em relação ao inglês. E, portanto, devido a um maior número de formas flexionadas por lema, existe um maior número de tokens e, consequentemente, menos tokens em comum entre os documentos em espanhol. Ao analisarmos as Figuras 4 e 6, ambas as *box plots* do subcorpo *i_es_od* resultam bastante similar às do *i_en_td* caso haja um valor médio de tokens maior por documento. Com a exceção do *whisker* mais longo na Figura 5, os valores do SCC também apresentam distribuições, médias e desvios padrão bastante similares quando comparados com o subcorpo *i_en_td* (veja-se Figura 1).

Apesar do subcorpo alemão *i_de_od* ter mais *tokens* e menos *types* (21,4k e 199,8k, respetivamente) quando comparado com o *i_es_od* (13k types e 207,3k tokens), o seu rácio $\frac{types}{tokens}$ não varia muito entre eles (0,049 contra 0.063, para mais detalhes veja-se Tabela 3). O mesmo ocorre com os valores do NTC, do SCC e do χ^2 (Figuras 1, 2 e 3). Por exemplo, o NTC entre os documentos, em média, para o subcorpo *i_es_od* é de 31,97 com um σ=23,48, contra uma av=43,21 e um σ=33,52 para o subcorpo *i_de_od*. Além disso, a média e o desvio padrão do seu SCC e χ^2 são ainda mais expressivos (ou seja, SCC=$\{$av=0,415 e σ=0,07$\}$ para o *i_es_od* vs. SCC=$\{$av=0,427$\}$ e σ=0,065 para o *i_de_od* e χ^2=$\{$av=40,922; σ=38,212$\}$ para o *i_es_od* vs. χ^2=$\{$av=48,235; σ=45,301$\}$ para o *i_de_od*).

Como podemos observar nas Figuras 4, 5 e 6, a média de valores por documento para ambos os subcorpos *i_es_td* e *i_de_td* são ligeiramente diferentes dos valores apresentados nas *box plots* do subcorpo *i_en_td*. Além do reduzido NTC por documento, os desvios padrão do χ^2 resultarem maiores que as suas médias (*i_es_td*=$\{$av=13,40; σ=18,95$\}$ e *i_de_td*=$\{$av=2,771; σ=2,883$\}$), e a expressiva variabilidade dentro e fora do IQR do SCC no subcorpo *i_es_td* indiciam uma certa inconsistência nos dados. Esta instabilidade poderá ser explicada pelo reduzido número de types (*i_es_td*=3,4k e *i_ed_td*=5,5k) e tokens (*i_es_td*=16,4k e *i_de_td*=26,8k) e pelo seu rácio $\frac{types}{tokens}$ de 0,207 e 0,205, respetivamente (Tabela 3). Como referido por Baker (2006), a análise do rácio $\frac{types}{tokens}$ torna-se útil quando estamos perante subcorpos de tamanho reduzido. Assim, é bastante interessante observar que estes dois subcorpos só têm em média 607 e 246 tokens (*i_es_td*=$\frac{16400}{27} \approx 607$ e *i_de_td*=$\frac{26800}{109} \approx 246$), e, 126 e 50 types por documento (*i_es_td*=$\frac{3400}{27} \approx 126$ e *i_de_td*=$\frac{5500}{109} \approx 50$), o que os converte numa excelente prova de conceito. Quando comparados com os baixos rácios dos demais subcorpos (ver Tabela 3), - mesmo para este tipo de corpos - estes valores podem muito bem serem considerados elevados. Deste modo, podemos concluir que o elevado rácio sugere que estamos perante uma forma mais diversificada do uso da linguagem, o que consequentemente também pode explicar os baixos valores no NTC e do χ^2 para estes dois subcorpos. Por outro lado, um rácio baixo também pode indicar um grande número de repetições (uma mesma palavra ocorrendo uma e outra vez), o que pode implicar que estamos perante um domínio bastante especializado. Apesar do elevado valor do SCC, os dados são assimétricos e variáveis (veja-se a grande amplitude inter-quartis na Figura 5). Isso acontece porque a maioria das entidades comuns ocorrem poucas vezes nos documentos e, consequentemente, estas posicionam-se próximas umas das outras nas listas de ranking, o que depois resulta em elevados valores no SCC, principalmente por causa da sua influência no numerador da fórmula (ver equação 1).

Depois de analisados os vários subcorpos, o próximo passo passou por entender como os documentos traduzidos afetariam a similaridade interna quando adicionados aos subcorpos originais correspondentes. Para esse fim, realizamos várias experiências adicionando diferentes percentagens de documentos traduzidos, selecionados aleatoriamente, aos subcorpos originais. Mais precisamente, começamos por adicionar 10%, 20%, 30% e por fim 100%[8] dos documentos aos subcorpos

[8] O número de documentos correspondentes a estas percentagens podem ser inferidas a partir da Tabela 3.

originais. As Figuras 13, 14 e 15 apresentam os valores médios por documento para cada uma das diferentes percentagens. Como esperado, quanto mais documentos são adicionados menor é o NTC (veja-se Figura 13). No entanto, é necessária uma análise mais profunda dos resultados obtidos.

Embora o NTC para o espanhol seja menor quando 100% dos documentos traduzidos são adicionados ao subcorpo original, resultando em ≈9.3% menos tokens comuns por documentos, a queda em si não é muito significativa. Na verdade, o valor médio de tokens por documento aumenta ≈1.19% e ≈1.22% quando adicionados 20% e 30% dos documentos traduzidos, respetivamente. A reduzida variação nos valores do SCC e χ^2 também corrobora este facto (veja-se Figuras 14 e 15, respetivamente). O mesmo fenómeno pode-se observar para o inglês quando são adicionados os documentos traduzidos. O subcorpo original tem uma av=163,70 tokens e quando 10%, 20%, 30% e 100% dos documentos traduzidos são adicionados o NTC somente diminuiu ≈3.2%, ≈3.4%, ≈6.1% e ≈23.6%, respetivamente.

Deste modo, podemos inferir com base nos resultados estatísticos obtidos, que caso um subcorpo com mais documentos seja necessário para uma determinada tarefa em particular, os respetivos documentos originais e traduzidos em espanhol e inglês podem ser adicionados sem que a sua similaridade interna seja gravemente comprometida. Mesmo que esta junção signifique que hajam alguns documentos ruidosos dentro dos novos subcorpos, particularmente para o espanhol esta união representa um aumento no número de documentos de ≈12% e, a uma perda de somente ≈9.3% no seu grau de similaridade interno. Apesar de uma diminuição de ≈23,6% no NTC para o inglês, o aumento no número de documentos é mais significativa que para o espanhol, mais precisamente de ≈39.7%.

Relativamente ao alemão, a união dos seus subcorpos resulta numa diminuição abrupta de ≈53.4% no grau interno de similaridade. Este facto é bem visível nas Figuras 13 e 15, o que nos leva a ser ainda mais cautelosos em relação à junção dos seus dois subcorpos.

Dado os resultados analisados até ao momento podemos afirmar, de um ponto de vista teórico e estatístico, que os subcorpos *i_en_od*, *i_en_td* e *i_de_od* agregam documentos com um elevado grau de similaridade. E, pelo contrário, o mesmo não se pode afirmar para os subcorpos *i_es_od*, *i_es_td* and *i_de_td*. A segunda conclusão a retirar dos dados analisados é que se fosse

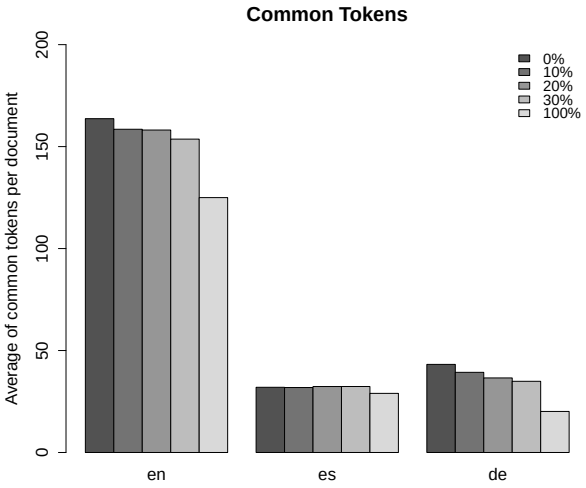


Figura 13: NTC.

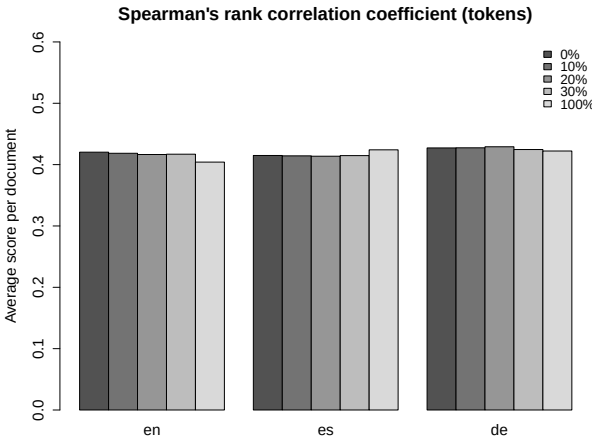


Figura 14: SCC.

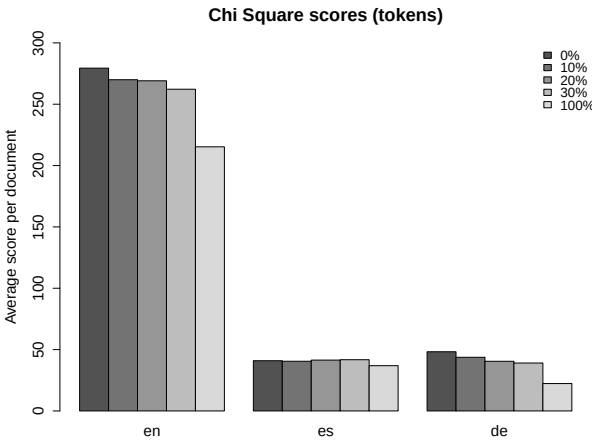


Figura 15: χ^2.

necessário um subcorpo especializado maior para o espanhol e/ou inglês, as evidências estatísticas mostram que ambos os seus subcorpos, originais e traduzidos, poderiam ser agregados sem que diminuísse drasticamente o seu grau de similaridade interno - especialmente para o espanhol em que a queda seria de apenas ≈9.3%.

Contudo, é aconselhável que qualquer tipo de trabalho de investigação seja feito no subcorpo original e, somente em casos que este não seja suficientemente grande para a tarefa em questão é que se deve prosseguir com a fusão com o respetivo subcorpo traduzido.

6.2 Compilação Manual vs. Semi-automática

Esta secção tem como objetivo comparar os subcorpos compilados manualmente com os corpos compilados semi-automaticamente pelo BootCaT (ver secção 4 para mais informação sobre os diversos subcorpos). Como não existem documentos traduzidos em italiano, decidiu-se realizar as seguintes experiências apenas usando os subcorpos originais (ou seja, usando os subcorpos *i_en_od*, *i_es_od*, *i_de_od* e *i_it_od* - ver Tabela 3). Em primeiro lugar foi feita uma comparação estatística entre os dois tipos de subcorpos de modo a compreender como a sua similaridade interna difere entre si. Em seguida, analisámos se a junção dos documentos compilado semi-automaticamente com o documentos originais comprometem o grau de similaridade interno dos mesmos.

De um modo semelhante ao que foi feito na secção anterior, as Figuras 16, 17 e 18 colocam lado a lado os valores médios por documento para as várias línguas (inglês, espanhol, alemão e italiano). A primeira observação que podemos fazer sobre a Figura 16 é a surpreendente diferença no NTC entre os documentos originais e os compilados semi-automaticamente. Por exemplo veja-se o NTC médio para o subcorpo *i_en_od* de 163,70 com um σ=83,89 quando comparado com o *bc_en* que apenas tem uma av=43.28 com um σ=56.97, ou seja, ≈74% menos tokens em comum por documento em média. De facto a diferença para o italiano é ainda maior, ≈91% menos tokens em comum por documento em média para sermos mais precisos (*i_it_od*={av=101,08; σ=55,71} e *bc_it*={av=9,26; σ=10,46}). Estes resultados podem ser corroborados pela variação dos valores do SCC e pelos baixos valores do χ^2 resultantes para o *bc_en* e para o *bc_it* quando comparados com os subcorpos *i_en_od* e *i_it_od*, respetivamente (Figuras 17 e 18). Contudo, note-se que o subcorpo *bc_en* tem vários *outliers* por cima do máximo, o que significa que estes documentos têm um elevado grau de similaridade com os do subcorpo *i_en_od* e, portanto, devem ser cuidadosamente analisados pela pessoa responsável pela manutenção do corpo.

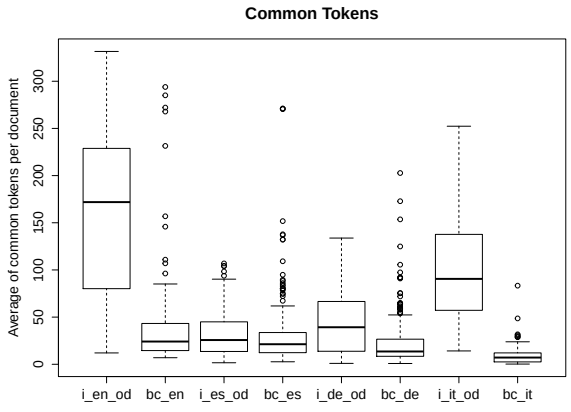


Figura 16: NTC.

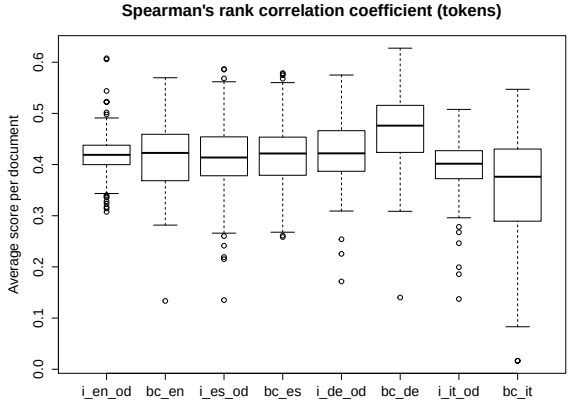


Figura 17: SCC.

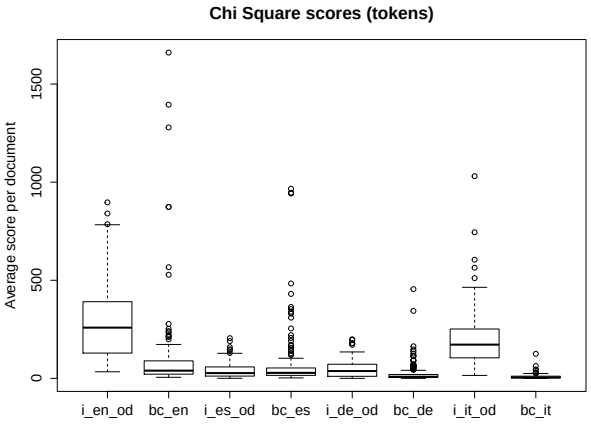


Figura 18: χ^2.

Relativamente ao subcorpo *bc_de*, este tem ≈22% menos tokens comuns por documento em média quando comparado com o subcorpo *i_de_od* (*i_de_od*={av=43,21; σ=33,52} e *bc_de*={av=23,06; σ=26,68}). Apesar desta diferença de 22% entre os dois subcorpos em alemão, não devemos rejeitar a hipótese de que estes dois subcorpos não podem ser unidos sem diminuir drasticamente o grau de similaridade interno - no entanto, é necessária uma análise mais profunda, como veremos mais

tarde nesta secção. Em relação aos subcorpos em espanhol, estes, à primeira vista, parecem conter documentos com um grau de similaridade idêntico, pois as suas médias e desvios padrão não diferem muito entre eles (*i_es_od*={av=31,97; σ=23,48} e *bc_es*={av=31,38; σ=36,51}). Além do mais, os valores do SCC e χ^2 também parecem confirmar esta hipótese (veja-se as Figuras 17 e 18).

Em suma, por um lado, os valores médios das MSD apresentados nas Figuras 16, 17 e 18 oferecem fortes evidências de que os subcorpos compilados manualmente e os compilados semi-automaticamente para o inglês e italiano não têm muito em comum. Por outro lado, as MSD sugerem que os subcorpos alemão e, principalmente os subcorpos espanhóis, partilham um elevado grau de similaridade entre os seus subcorpos e, portanto, a sua união pode ser considerada caso necessário. Para pôr à prova estes indícios, aleatoriamente selecionámos e adicionámos diferentes percentagens de documentos compilados semi-automaticamente aos subcorpos originais. A nossa hipótese é que os valores médios das MSD diminuam quanto mais documentos semi-automaticamente compilados são adicionados. Com base nos resultados anteriores, é esperada uma queda drástica para o inglês e italiano e uma queda mais suave para o alemão e, particularmente, para o espanhol.

As Figuras 19, 20 e 21 apresentam os valores médios por documento quando adicionadas diferentes percentagens de documentos semi-automaticamente compilados aos subcorpos originais. De modo a entendermos como o grau interno de similaridade varia, foram aleatoriamente selecionados e incrementalmente adicionados conjuntos de 10% aos subcorpos originais. Acima de tudo o que é importante analisar nas Figuras 19, 20 e 21 é o seguinte: i) os valores médios iniciais, ou seja os valores dos subcorpos compilados manualmente (0%); ii) como estes valores variam quando mais documentos são adicionados (de 10% a 100%); iii) e comparar o valor inicial com o valor final, ou seja quando a totalidade dos documentos semi-automáticos é adicionada ao subcorpo original (0% e 100%). Já anteriormente, quando colocámos as Figuras 16, 17 e 18 lado a lado, deu para ter uma ideia sobre o que aconteceria quando fosse feita esta união dos dois tipos de subcorpos e, de facto as Figuras 19 e 21 vêm corroborar a nossa tese inicial. Como podemos ver na Figura 19, quanto mais conjuntos de documentos são adicionados, menor é o NTC para as quatro línguas de trabalho.

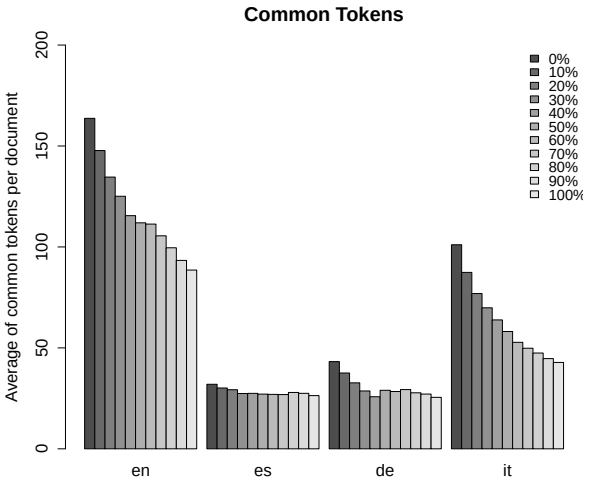


Figura 19: NTC.

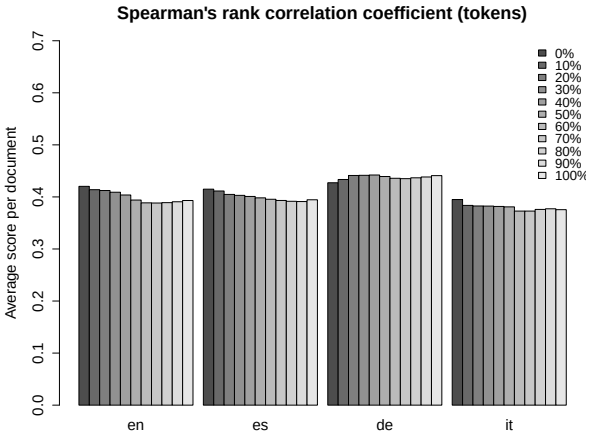


Figura 20: SCC.

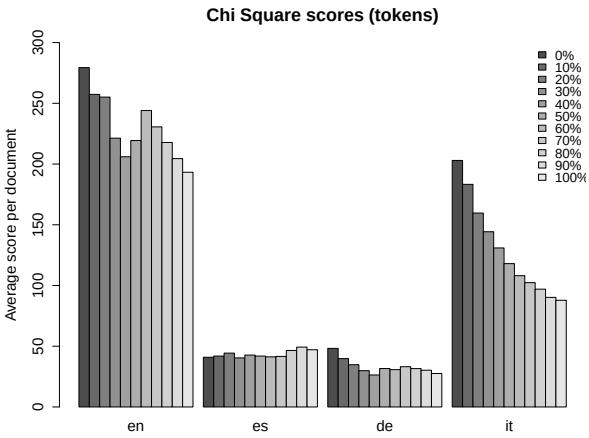


Figura 21: χ^2.

Como mencionado anteriormente, o NTC por documento para o subcorpo *i_en_od* é, em média, de 163,70. Contudo, quando o *bc_en* é adicionado - o que significa um aumento de ≈73.5% em termos de tamanho - o NTC diminui para

quase metade (ou seja, há uma diminuição de ≈46%: {*i_en_od* + *bc_en*}={av=88.55}). Para o italiano a redução do NTC é ainda mais acentuada, mais precisamente de ≈58% ({*i_it_od* + *bc_it*}={av=42.79}, enquanto que o aumento no número de documentos é de ≈81.3%). E, o alemão segue a mesma tendência com uma redução no NTC de ≈41%, contudo é necessário ter em conta que esta união representa um aumento no número de documentos de ≈183.3%. Os valores do χ^2 também apontam na mesma direção, ou seja, os valores do χ^2 diminuem em ≈31%, ≈57% e ≈43% para os subcorpos {*i_en_od* + *bc_en*}, {*i_it_od* + *bc_it*} e {*i_de_od* + *bc_de*}, respetivamente. Um fenómeno semelhante ocorre com o espanhol, observe-se a Figura 16. Contudo, e apesar da diminuição do NTC em ≈17% para o espanhol quando este sofre um aumento de ≈103.8% no número de documentos, o grau de similaridade interno parece estabilizar assim que o primeiro conjunto de documentos é adicionado, o que poderá significar que o subcorpo *bc_es* segue uma distribuição normal em termos de conteúdo, neste caso no NTC por documento. Em relação aos valores do χ^2, este sofre um aumento de ≈15%, o que mostra indícios de um aumento da similaridade interna.

De forma semelhante à conclusão retirada na secção 6.1 (quando comparámos os subcorpos originais com os traduzidos), os valores do NTC e os valores χ^2 das Figuras 19 e 21, assim como os resultados observados nas Figuras 16, 17 e 18, leva-nos a concluir que caso seja necessário um maior subcorpo especializado para o espanhol a união entre os textos originais e os compilados semi-automaticamente pode ser realizada sem que o grau interno de similaridade seja drasticamente comprometido. Ou, pelo menos, é mais aconselhável sugerir esta união do que a união dos subcorpos do italiano, do alemão ou mesmo do inglês. Embora, em geral, os valores do SCC diminuam para três das quatro línguas, estes, no entanto, não são suficientemente explícitos para nos permitir tirar uma conclusão sólida sobre os mesmos (veja-se Figura 20).

6.3 Discussão

Depois de apresentados todos os resultados estatísticos é hora de seguir em frente e analisar o problema de uma perspetiva diferente e centrarmo-nos sobre a seguinte questão: "*Devemos sempre confiar nas ferramentas semi-automáticas para compilar corpos comparáveis especializados?*". A questão em si é simples, mas como foi demonstrado nas secções anteriores, a resposta não é trivial. Por um lado, podemos assumir que as ferramentas de compilação semi-automáticas têm uma abrangência maior quando comparadas com a compilação manual, pois estas são capazes de compilar mais documentos do que um humano no mesmo espaço de tempo. Contudo, a sua precisão não é tão elevada como a de um humano - embora esta ideia seja discutível, o humano é quem tem a última palavra a dizer e, consequentemente, aquele que julga se os documentos devem pertencer ao corpo ou não. Porém, também podemos afirmar que a compilação manual nem sempre é viável, uma vez que é muito demorada e exige um grande esforço intelectual. Na verdade é que derivado à enorme quantidade de variáveis envolvidas no processo de compilação, tais como o domínio, as línguas de trabalho, os motores de busca utilizados, entre outros, que não se pode afirmar que exista uma resposta simples para a questão anterior. Por exemplo, cada motor de busca utiliza um método de indexação diferente para armazenar e encontrar páginas na rede, o que significa que diferentes motores de busca devolvem diferentes resultados. De volta à questão, e com base nos nossos resultados, o que podemos afirmar é que as ferramentas de compilação semi-automáticas podem-nos ajudar a impulsionar o processo de compilação. E, embora algumas fases do processo possam ser semi-automatizadas, estas ferramentas não funcionam corretamente sem a intervenção humana. Contudo, devemos ter sempre muito cuidado ao compilar corpos comparáveis em geral e corpos comparáveis especializados em particular, não só durante o processo inicial de *design*, mas também na última instância do processo de compilação, ou seja, ao analisar e filtrar os documentos compilados que devem fazer parte do corpo. E, é precisamente nesta etapa do processo onde a metodologia proposta neste trabalho se encaixa, podendo não só ser usada para ter uma ideia sobre os documentos em mãos, mas também para comparar diferentes conjuntos de documentos, e classificar os mesmos de acordo com o seu grau de similaridade. Deste modo, a pessoa em cargo da compilação poderá usar esta metodologia como uma ferramenta extra para a ajudar a descrever um corpo e até mesmo para decidir se um determinado documento ou conjunto de documentos devem fazer parte do mesmo ou não.

7 Conclusão

Neste artigo descrevemos uma metodologia simples, contudo eficiente, capaz de medir o grau de similaridade no contexto de corpos comparáveis. A metodologia apresentada reúne vários métodos de diferentes áreas do conhecimento com a finalidade de descrever, medir e classificar documentos com base no conteúdo partilhado entre eles. De modo a provar a sua eficácia foram realizadas várias experiências com três diferentes Medidas de Similaridade Distribucional (MSD).

Resumidamente, a primeira parte deste trabalho focou-se na análise dos diversos subcorpos compilados manualmente e as principais conclusões foram as seguintes: i) foram obtidos resultados semelhantes utilizando diferentes parâmetros de entrada para as várias MSD; ii) os documentos originais contêm um maior número de entidades comuns quando comparados com os traduzidos; e iii) as MSD sugerem que os subcorpos em inglês e italiano originais são compostos por documentos com um maior grau de similaridade em comparação com os restantes subcorpos analisados neste trabalho. O passo seguinte passou por demonstrar como os documentos traduzidos afetariam o grau de similaridade interno nos vários subcorpos originais quando unidos. Embora o grau de similaridade tenha reduzido drasticamente, ≈53,4% para o alemão após a fusão, o subcorpo espanhol e inglês diminuiu apenas ≈23,6% e ≈9,3%, respetivamente. Deste modo, demos por concluída a primeira parte deste trabalho afirmando que, caso fosse necessário um subcorpo especializado maior para o espanhol ou inglês, as MSD demonstraram que a união entre o subcorpo original e o subcorpo traduzido poderia ser realizada sem que se reduza drasticamente o seu grau interno de similaridade.

A segunda parte deste trabalho focou-se na comparação entre os documentos compilados manualmente e os documentos compilados semi-automaticamente. Mais uma vez começámos por realizar uma análise estatístico-descritiva entre os dois tipos de documentos de modo a obter uma ideia geral de como a similaridade média interna diferia entre eles. Como resultado, observou-se que os subcorpos compilados manualmente continham documentos com um maior grau de similaridade quando comparados com os correspondentes subcorpos compilados semi-automaticamente. Especialmente para o inglês e italiano, observamos que a diferença entre a média no número de entidades comuns era muito elevada, para sermos mais precisos, ≈74% e ≈91% menos entidades comuns, respetivamente. Estes valores já nos dão uma ideia sobre o que ocorreria quando uníssemos os subcorpos compilados manualmente com os semi-automáticos. De modo a demonstrar a sua veracidade, juntámos os vários subcorpos e as MSD demonstraram uma queda drástica em termos de similaridade interna. Mais precisamente, foi observada uma queda muito acentuada, na ordem dos 41%, 46% e 58% para o alemão, inglês e italiano, respetivamente, e uma queda não tão abrupta de ≈17% para o espanhol. Com estes resultados, concluímos que caso fosse necessário um subcorpo especializado maior para o espanhol, esta união deveria ser ponderada. Pois, se por um lado a similaridade interna caíra 17%, por outro, esta união aumentaria o número de documentos em ≈109.8%.

Como observação final, concluímos que as várias MSD podem ser consideradas uma ferramenta muito útil e versátil para descrever corpos comparáveis, o que na nossa opinião ajudaria em muito aqueles que compilam manualmente ou semi-automaticamente corpos a partir da Internet nas mais diversas línguas europeias. De facto, este trabalho provou que as MSD não só podem ser utilizadas para obter uma ideia sobre o corpo em mãos, mas também para medir, comparar e classificar diferentes conjuntos de documentos de acordo com o seu grau de similaridade e assim ajudar os investigadores a decidir se um determinado documento ou conjunto de documentos devem fazer parte de um dado corpo ou não.

Agradecimentos

Gostaríamos de agradecer à Bárbara Furtado e ao João Miguel Franco pelas correções ortográficas e gramaticais no artigo.

Hernani Costa é apoiado pela bolsa n. 317471 da REA do People Programme (Marie Curie Actions) da European Union's Framework Programme (FP7/2007-2013).

Este trabalho também é parcialmente apoiado pelo projeto de inovação para a educação TRADICOR (PIE 13-054, 2014-2015); pelo projeto de inovação para a educação NOVATIC (PIE 15-145, 2015-2017); o projeto de I&D INTELITERM (ref. n. FFI2012-38881, 2012-2015); o projeto de I&D LATEST (Ref: 327197-FP7-PEOPLE-2012-IEF); e o projeto de I&D TERMITUR (ref. n. HUM2754, 2014-2017).

Referências

Anthony, Laurence. 2014. AntConc (Version 3.4.3) Machintosh OS X. Waseda University. Tokyo, Japan. `http://www.laurenceanthony.net`.

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. Bloomsbury Discourse. Bloomsbury Academic.

Barbaresi, Adrien. 2014. Finding Viable Seed URLs for Web Corpora: A Scouting Approach and Comparative Study of Available Sources. Em *9th Web as Corpus Workshop (WaC-9), 14th Conf. of the European Chapter of the Association for Computational Linguistics*, pp. 1–8, Gothenburg, Sweden.

Barbaresi, Adrien. 2015. Challenges in the linguistic exploitation of specialized republishable web corpora. Em *RESAW Conf. 2015*, pp. 53–56, Aarhus, Denmark. Short paper talk.

Baroni, Marco e Silvia Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. Em *4th Int. Conf. on Language Resources and Evaluation*, LREC'04, pp. 1313–1316.

Baroni, Marco, Adam Kilgarriff, Jan Pomikálek, e Pavel Rychlý. 2006. WebBootCaT: instant domain-specific corpora to support human translators. Em *11th Annual Conf. of the European Association for Machine Translation*, EAMT'06, pp. 247–252, Oslo, Norway. The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway).

Bowker, Lynne e Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Corpas Pastor, Gloria. 2001. Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, 5(1):155–184.

Corpas Pastor, Gloria e Míriam Seghiri. 2009. Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). Em A. Beeby, P.R. Inés, e P. Sánchez-Gijón, editores, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library. John Benjamins Publishing Company, capítulo 5, pp. 75–107.

Costa, Hernani. 2010. Automatic Extraction and Validation of Lexical Ontologies from text. Tese de Mestrado, University of Coimbra, Faculty of Sciences and Technology, Department of Informatics Engineering, Coimbra, Portugal, September, 2010.

Costa, Hernani. 2015. Assessing Comparable Corpora through Distributional Similarity Measures. Em *EXPERT Scientific and Technological Workshop*, pp. 23–32, Malaga, Spain, June, 2015.

Costa, Hernani, Hanna Béchara, Shiva Taslimipoor, Rohit Gupta, Constantin Orasan, Gloria Corpas Pastor, e Ruslan Mitkov. 2015. MiniExperts: An SVM approach for Measuring Semantic Textual Similarity. Em *9th Int. Workshop on Semantic Evaluation*, SemEval'15, pp. 96–101, Denver, Colorado, June, 2015. ACL.

Costa, Hernani, Gloria Corpas Pastor, e Ruslan Mitkov. 2015. Measuring the Relatedness between Documents in Comparable Corpora. Em *11th Int. Conf. on Terminology and Artificial Intelligence*, TIA'15, pp. 29–37, Granada, Spain, November, 2015.

Costa, Hernani, Hugo Gonçalo Oliveira, e Paulo Gomes. 2010. The Impact of Distributional Metrics in the Quality of Relational Triples. Em *19th European Conf. on Artificial Intelligence, Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, ECAI'10, pp. 23–29, Lisbon, Portugal, August, 2010.

Costa, Hernani, Hugo Gonçalo Oliveira, e Paulo Gomes. 2011. Using the Web to Validate Lexico-Semantic Relations. Em *15th Portuguese Conf. on Artificial Intelligence*, volume 7026 of *EPIA'11*, pp. 597–609, Lisbon, Portugal, October, 2011. Springer.

de Groc, Clement. 2011. Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. Em *IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology*, volume 1 of *WI-IAT'11*, pp. 497–498, Lyon, France, August, 2011. IEEE Computer Society.

EAGLES. 1996. Preliminary Recommendations on Corpus Typology. Relatório técnico, EAGLES Document EAG-TCWG-CTYP/P., May, 1996. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

Grishman, Ralph. 1997. Information Extraction: Techniques and Challenges. Em *Int.*

Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, SCIE'97, pp. 10–27, London, UK. Springer.

Gutiérrez Florido, Rut, Gloria Corpas Pastor, e Miriam Seghiri. 2013. Using semi-automatic compiled corpora for medical terminology and vocabulary building in the healthcare domain. Em *Workshop on Optimizing Understanding in Multilingual Hospital Encounters, 10th Int. Conf. on Terminology and Artificial Intelligence (TIA'13)*, Paris, France.

H. El-Khalili, Nuha, Bassam Haddad, e Haya El-Ghalayini. 2015. Language Engineering for Creating Relevance Corpus. *Int. Journal of Software Engineering and Its Applications*, 9(2):107–116.

Harris, Zelig. 1970. Distributional Structure. Em *Papers in Structural and Transformational Linguistics*. D. Reidel Publishing Company, Dordrecht, Holland, pp. 775–794.

Ibrahimov, Oktay, Ishwar Sethi, e Nevenka Dimitrova. 2002. The Performance Analysis of a Chi-square Similarity Measure for Topic Related Clustering of Noisy Transcripts. Em *16th Int. Conf. on Pattern Recognition*, volume 4, pp. 285–288. IEEE Computer Society.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, e Vít Suchomel. 2014. Finding Terms in Corpora for Many Languages with the Sketch Engine. Em *Demonstrations at the 14th Conf. of the European Chapter of the Association for Computational Linguistics*, pp. 53–56, Gothenburg, Sweden. ACL.

Kilgarriff, Adam. 2001. Comparing Corpora. *Int. Journal of Corpus Linguistics*, 6(1):97–133.

Maia, Belinda. 2003. What are comparable corpora? Em Silvia Hansen-Schirra e Stella Neumann, editores, *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*, pp. 27–34, Lancaster, UK, March, 2003.

Rayson, Paul, Geoffrey Leech, e Mary Hodges. 1997. Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. *Int. Journal of Corpus Linguistics*, 2(1):133–152.

Salton, Gerard e Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Schmid, Helmut. 1995. Improvements In Part-of-Speech Tagging With an Application To German. Em *ACL SIGDAT-Workshop*, pp. 47–50, Dublin, Ireland.

Singhal, Amit. 2001. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–42.

Costa et al. (2017)

Costa, H. and Corpas Pastor, G. and Durán Muñoz, I. (2017) **Assessing Terminology Management Systems for Interpreters**. In Corpas Pastor, Gloria and Durán Muñoz, Isabel, *Trends in E-Tools and Resources for Translators and Interpreters*, volume 45, pages 57-84, Brill.

Assessing Terminology Management Systems for Interpreters

Hernani Costa
University of Malaga, Malaga, Spain
hercos@uma.es

Gloria Corpas Pastor
University of Malaga, Malaga, Spain
g.corpas@uma.es

Isabel Dúran Muñoz
University of Alcalá de Henares, Madrid, Spain
iduran@uma.es

Abstract

This paper aims at describing and comparing current Terminology Management Systems (TMS) with a view to establishing a set of features in order to assess the extent to which terminology tools meet the specific needs of interpreters. As in translation, domain-specific terminology becomes a cornerstone in interpreting when consistency and accuracy are at stake. Hence, an efficient use and management of terminology will enhance interpreting results. As a matter of fact, interpreters have limited time to prepare for new topics and they have to carry out searches and preparation prior to an interpretation and have it accessible during the interpreting service. Fortunately, there is an ever-growing number of applications capable of assisting interpreters before and during an interpretation service, even though they are still few if compared to those devoted to translators. Although these tools appear to be quite similar, they provide different kind of features which result in different degrees of usefulness, as it can be observed in the last section of this paper.

Keywords

interpreter's needs, interpretation service, interpreting, language technology, terminology management systems, preparation phase.

1 Introduction

Interpreting can be distinguished from other types of translation processes by its immediacy. Following Pöchhacker (2007, p.10), "Interpreting is performed here and now for the benefit of people who want to engage in communication across barriers of language and culture.". Currently, there is no universally accepted classification of interpreting modes, since authors and interpreting institutions, such as ITI or DG Interpretation at the European Union, propose their own classifications. However, the most frequent interpreting modes encountered in the literature and offered by company services are based on the three following criteria: timing/delay of relaying the translated message, direction of interpreting and setting/purpose of the interaction.

Depending on the timing/delay of relaying the translated message, the main categories of interpreting are *simultaneous interpreting* and *consecutive interpreting*. Simultaneous interpreting is defined as a translated message that is given at roughly the same time that the source message is produced. In consecutive interpreting the interpreter waits until the speaker has finished before beginning the interpretation and takes notes in the meantime.

Depending on the direction of interpreting, we can distinguish *unidirectional interpreting* and *bi-lateral* or *bi-directional interpreting*. Unidirectional interpreting occurs in situations in which the message is conveyed to a passive audience, and bi-lateral or bi-directional interpreting happens when the interpreter mediates/facilitates communication/dialogue between two parties (also called *liaison interpreting*).

Depending on the setting/purpose of the interaction, we can distinguish:

- *Conference*: Simultaneous interpreting at international conferences and formal meetings, with interpreters working in pairs;
- *Business*: Interpreting at smaller or less formal company meetings, factory visits, exhibitions, product launches, government meetings and accompanying delegations etc.;
- *Police and court*: Interpreting for the police and courts, the probation service, solicitors,

arbitrations and tribunals etc.;

- *Community*: Interpreting for individuals and organisations such as the NHS, social services in matters of health and welfare, the local government, not-for-profit or charitable organisations and at community events.

Teleinterpreting (also *remote interpreting*) is an important modality of interpreting provided by a remote or offsite interpreter via telephone (over the phone interpreting) or via video (video remote interpreting). This is usually done in consecutive mode, but simultaneous interpreting is possible depending on the capabilities of the telecommunication technology used. Other modalities are whispered interpreting, sign language interpreting, sight interpreting and others.

Despite the different modes of interpretation, it is hardly impossible for interpreters to collect the relevant specialised information during the interpretation service itself. They frequently face different settings and specialised fields in their interpretation services and yet they always need to provide excellent results. They might be called to work for specialists that share a background knowledge that is totally or partially unknown to laypersons and/or outsiders (Will, 2007). When interpreters lack the necessary background knowledge or experience, they usually need to perform extensive searches for specialised knowledge and terminology in a very efficient way in order to supply this deficit and acquire the required information. In this sense, interpreters are required to find the relevant information for their service prior to interpretation and have it accessible during the process.

As it is well known, interpreting is an extremely strenuous task, since it involves much effort in terms of decoding, memorising and encoding a message (Tripepi Winteringham, 2010, p.88). Therefore, interpreters should, as other professionals do, benefit from the development of technology, which will bring about a considerable improvement of their working conditions (Costa, Corpas Pastor, and Durán Muñoz, 2014a). Where language technologies are concerned, advances have been observed due to the confluence of telecommunications and digital data processing systems in the last decades (Pöchhacker, 2007, p.168). However, language technology developments need more systematic research. To date, a limited number of studies have focused on the needs of interpreting technology (Moser-Mercer, 1992; Berber, 2010; Braun, 2006; Kalina, 2010), to Computer-Assisted Interpreter Training (CAIT) (Gran, Carabelli, and Merlini, 2002; de Manuel Jerez, 2003; Blasco Mayor, 2005; Sandrelli and de Manuel Jerez, 2007) or on Computer-Assisted Interpreting (CAI) tools (Kelly, 2009; Tripepi Winteringham, 2010; Costa, Corpas Pastor, and Durán Muñoz, 2014b; Costa, Corpas Pastor, and Durán Muñoz, 2014a; Costa, Corpas Pastor, and Durán Muñoz, 2015; Zhang, 2016; Fantinuoli, 2016). Although some interpreters have shown some degree of reluctance to use language technologies in their profession (see Berber (2010)), it is clear that CAI tools represent an important advance in the field of interpretation and thus in the multilingual communication context. Nevertheless, the solutions tailored to interpreters' needs are few and still far behind (Costa, Corpas Pastor, and Durán Muñoz, 2014b; Costa, Corpas Pastor, and Durán Muñoz, 2014a).

In this paper, we aim to shed some light on a specific type of technology targeting interpreters – Terminology Management Systems (TMS) – and to carry out a comparative analysis of several of those tools in order to assess their relevance.

2 Interpreter's Terminology Needs

The potentialities of computers for improving interpreters' working conditions was pointed out by Gile (1987) long time ago. However, very little progress has been made so far. Costa, Corpas Pastor, and Durán Muñoz (2014b) offer a tentative catalogue of current language technologies for interpreters, divided into terminology tools for interpreters, note-taking applications for consecutive interpreting, applications for voice recording and training tools. This paper focuses exclusively on terminology tools for interpreters with a view to performing a user evaluation.

As a rule, most interpreters seem to be unaware of the opportunities offered by language technologies. As far as terminology is concerned, interpreters continue to store information and terminology on scraps of paper or excel spreadsheets, while the use of technologies and terminology management tools is still very low. A study conducted by Moser-Mercer (1992, p.507) rejected the assumption that "interpreters' needs are identical to those of translators and terminologists" and intended to "survey how conference interpreters handle terminology documentation and document control and to offer some guidelines as to

the interpretation-specific software tools for terminology and documentation management". The results of this study includes some key findings, such as the conclusion that most of the respondents were interested in exchanging terminological information and that they were open to using computers in their profession. According to these findings, Moser-Mercer (1992) highlighted that "software developers targeting the conference interpreting market must provide a tool that meets the specific needs of the interpreters and not just market translation tools" (ibid:511). More recent studies have also studied interpreters' current needs and practices regarding terminology management (Rodríguez and Schnell, 2009; Bilgen, 2011), and they also share the same findings: interpreters require specific tools to meet their needs, which are different from translators and terminologists. According to a survey conducted by Bilgen (2011), 85% of respondents are open to using computers, yet conventional methods still prevail over the use of computerised methods of terminology management. The author observed that respondents had no or little experience with terminology management software, and those with some experience were most dissatisfied with the money and time they had to invest in them, and their overall experience was mediocre (ibid:66). Respondents indicated that their priorities were different from those identified in terminology literature in terms of terminological information stored, and the way in which term records are structured. This is an important aspect that differentiates the needs of interpreters and translators as regards definitions and contexts (Bilgen, 2011). Due to their working conditions, translators usually prefer to consult multiple definitions and contexts to find the best solution for the translation problem. On the contrary, interpreters will rarely have the time to go over multiple definitions, contexts, etc. to find the right one, and thus, they will need to store the most concise information to be able to consult it in the quickest and easiest way. Their responses in this survey also showed that the way they retrieve terminological information was context-specific, and that there was also a significant variation among individual interpreters. Flexibility is, therefore, of great importance to interpreters due to the variation of their context-specific terminology management practices, and on their individual preferences regarding the storage, organisation and retrieval of terminological information (ibid: 92). Rodríguez and Schnell (2009), after a thorough analysis of interpreters' needs and in order to meet their requirements as regards terminology management tools, propose the possibility of developing small databases that vary according to the area of speciality or according to the conference and client. These mini-databases would be multilingual and include an option allowing the interpreter to switch the source and target languages. This assumption is in line with the Function Theory (Bergenholtz and Tarp, 2003; Tarp, 2008) and electronic multifunctional dictionaries (Spohr, 2009), which both defend the need to elaborate terminological entries according to potential users. Rodríguez and Schnell (2009) recognise five features that would distinguish the interpreters' mini-databases from the terminology databases intended for translators:

- speed of consultation;
- intuitive navigation;
- possibility of updating the terminology record in the interpretation booth;
- considerable freedom to define the basic structure;
- multiple ways of filtering data.

Accordingly, they also suggest the abandonment of the usual terminology methodology if the intention is to provide interpreters with specific glossaries tailored to their needs. The authors propose the use of a semasiological and associative methodology instead of the onomasiological approach as the latter would slow down the interpretation process due to the extra cognitive effort required by onomasiological structures.

Bearing those features in mind, the next sections will describe and compare several TMS developed for or by interpreters to assess the extent to which these terminology tools meet the specific needs of the interpreters.

3 A Brief Survey of TMS

It is a well-known fact that terminology work is present in the whole process of preparation prior to an interpretation service. For example, interpreters become familiar with the subject field by searching for specialised documents, by extracting terms and looking for synonyms and hyperonyms, by finding and developing acronyms and abbreviations and by compiling a glossary. According to Rodríguez and Schnell (2009), interpreters tend to compile in-house glossaries tailored to their individual needs as the

main way to prepare the terminology of a given interpretation. As previous studies and surveys have shown, this terminology management carried out by interpreters is frequently done manually or with very little help of technology.

However, in the last decade a wealth of Terminology Management Systems (TMS) that interpreters could use to quickly compile, store, manage and search within glossaries have been developed. They can be typically used to prepare an interpretation, in consecutive interpreting or in a booth. Even though most of these TMS have not been specifically developed for interpreters but for translators, there are some of them that cater for the needs of both translators and interpreters (Durán Muñoz, 2012; Costa et al., 2016). Due to space constraints, only the TMS developed for interpreters that are currently available, together with some other TMS that can be useful in their interpreting tasks, are described in detail below.

3.1 Standalone TMS[1]

Intragloss[2] is a commercial Mac OS X software created specifically to help interpreters when preparing for an event by allowing them to manage glossaries. This application can be simply defined as a glossary and document management tool created to help the interpreter prepare, use and merge different glossaries with preparation documents, in more than 180 different languages. It permits to import and export glossaries from and to plain text, Microsoft Word and Excel formats. Every glossary imported to, or created in, is assigned to a domain glossary (considered the highest level of knowledge), which contains all the glossaries from the sub-areas of knowledge, named 'assignments'. The creation of an assignment glossary can be done in two different ways: either by extracting automatically all the terms from the domain glossary that appear in the imported documents, or by highlighting a term in the document, searching for it on search sites (such as online glossaries, terminology databases, dictionaries and general Web pages) and manually adding the new translated term to the assignment glossary. It is important to mention that the online search can be made within Intragloss. Another interesting feature is that Intraglosss allows users to copy assignment glossaries and assignment entries from one assignment to another. The domain glossary may be multilingual as it can include several bilingual assignment glossaries. By way of example, if there are two assignment glossaries English/French and Dutch/English, in the same domain, the domain glossary will be French/English/Dutch, i.e. multilingual. Finally, Intragloss also permits to manually add meta-information to each glossary entry.

In short, Intragloss is an intuitive and easy-to-use tool that facilitates the interpreters' terminology management process by producing glossaries (imported or created ad hoc), by searching on several websites simultaneously, by highlighting all the terms in the documents that appear in the domain glossary and by comparing different language versions of a document. However, it is currently platform dependent and only works on Mac OS X platforms.

InterpretBank[3] is a simple terminology and knowledge management software tool designed both for interpreters and translators using Windows and Android. It helps to manage, learn and look up glossaries and term-related information. Due to its modular architecture, it can be used to guide the interpreter during the entire workflow process, starting from the creation and management of multilingual glossaries (TermMode), passing through the study of these glossaries (MemoryMode), and finally allowing the interpreter to look up terms while in a booth (ConferenceMode). InterpretBank also has an Android version called InterpretBank Lite. This application is specifically designed to access bi- or trilingual glossaries previously created with the desktop version. It is useful when working as a consecutive, community or liaison interpreter, when a quick look up at the terminology list is necessary.

InterpretBank has a user-friendly, intuitive and easy-to-use interface. It allows us to import and export glossaries in different formats (Microsoft Word, Microsoft Excel, simple text files, Android and TMEX) and suggests translation candidates by taking advantage of online translation portal services, such as Wikipedia, MyMemory and Bing. However, it is platform-dependent (it only works on Windows and Android), does not handle documents (only glossaries) and requires a commercial license.

[1] The TMS are divided into three different categories: standalone, web-based and mobile TMS for the sake of clarification.

[2] `https://intragloss.com/`

[3] `www.interpretbank.de`

Interplex UE[4] is a user-friendly multilingual glossary management program that can be used easily and quickly in a booth while the interpreter is working. Instead of keeping isolated word lists, it allows to group all terms relating to a particular subject or field into multilingual glossaries that can be searched in an instant. This program permits to have several glossaries open at the same time, which is a very useful feature if the working domain is covered by more than one glossary. Similar to the previous analysed programs, Interplex UE also allows to import and export glossaries from and to Microsoft Word, Excel, and simple text files. Interplex UE runs on Windows; nevertheless, it has a simpler version for iOS devices, one named Interplex Lite, for iPhone and iPod Touch, and another named Interplex HD, for iPad. Both glossaries and multi-glossary searchers offer the functionality of viewing expressions in each of the defined languages.

In general, Interplex UE has a user-friendly interface and it is regularly updated. It allows to import and export glossaries from and to Microsoft Word and Excel formats. However, it is also platform dependent (only works on Windows and iOS), does not handle documents, only glossaries, and requires a commercial license.

SDL MultiTerm Desktop[5] is a commercial TMS developed for Windows that provides one solution to store and manage multilingual terminology. MultiTerm was first launched in 1990 by Trados GmbH but in 2005 the company was acquired by SDL , which renamed MultiTerm to SDL MultiTerm. Today, SDL MultiTerm is a terminology management tool commercialised by SDL[6] as a standalone application, which has been improved according to translators' needs. Alternatively, MultiTerm can be used within the SDL Trados Studio[7] as an integrated tool. As translators/interpreters can easily edit and add terminology within SDL Trados Studio, MultiTerm helps to improve the efficiency of the translation process and promotes high-quality translated content with real-time verification of multilingual terminology. This application is very complete because it allows to store an unlimited number of terms in a vast number of languages; imports and exports glossaries from and to different technology environments, such as Microsoft Excel, XML, TBX and several other proprietary formats; permits to manually add a variety of meta-data information, such as synonyms, context, definitions, associated project, part-of-speech tags, URLs, etc. Apart from the previous mentioned descriptive fields, MultiTerm also allows the user to insert illustrations for the terms in the terminology database (which can be stored either locally or, for collaborative purposes, in a remote server). This visual reference feature is very useful especially to interpreters and translators dealing with unfamiliar terms. Moreover, MultiTerm has an advanced search feature that permits to search not only the indexed terms but also in their descriptive fields, or create filters to make custom searches within specific fields, like language, definition, part-of-speech, etc. Nevertheless, the most interesting feature about MultiTerm is its concept-oriented feature, i.e. each entry in MultiTerm corresponds to a single concept, which can be described by different terms in both source and target language. This detail is very important because it allows the user to centralise and customise the terms with more information, such as different possible translations and their corresponding contexts.

In general, MultiTerm can be seen as an advanced multilingual TMS with an intuitive and easy-to-use interface. Although MultiTerm was originally designed for translators, it can also be used by interpreters. Its main advantage to interpreters, when compared with other terminology tools, is twofold: it allows users to add several translation terms in one entry and permits to customise a wide variety of descriptive fields, such as illustrations, associated projects, definitions, etc. However, it can only be used on Windows, does not handle documents and there is no demo version available.

AnyLexic[8] is an easy-to-use TMS developed for Windows with a simple and intuitive interface. It was not designed for any particular terminological requirement, instead it aims to help the interpreter prepare, use and manage different glossaries or dictionaries. AnyLexic can be described as a robust terminology management tool, as it enables users to easily create and manage multiple mono-, bi- or multilingual glossaries in any language and to import and export glossaries from and to Microsoft Excel, plain text and AnyLexic Exchange Format (AEF). In addition, each entry in the glossary can have multiple translation

[4] www.fourwillows.com

[5] www.sdl.com/cxc/language/terminology-management/multiterm/

[6] www.sdl.com

[7] www.sdl.com/products/sdl-trados-studio

[8] www.anylexic.com

equivalents in the target language along with notes. The search for records in the database allows users to combine different options, such as search for all source terms or translation candidates and associated notes. In addition, the search can be performed within one or multiple glossaries. Another interesting feature in AnyLexic is the way that records can be displayed using different templates with configurable text colour, background colour, font size and text format. Besides, it is possible to customise the template for displaying the records. With the purpose of simplifying the teamwork process, this tool has an additional option to exchange any glossary with other AnyLexic users by either using the AEF proprietary format or by accessing a remote glossary, a very useful feature for collaborative interpreting and/or translation projects.

In general, AnyLexic is an easy and convenient terminology database managing software for working with terminology, creating, editing and exchanging glossaries. However, it only works on Windows platforms and even though an evaluation version is available for 30 days, it requires a commercial license.

Lingo[9] is a commercial Windows terminology management tool designed to create and manage terminology databases, whether mono- or multilingual. It can import from and export to TMX and plain text. Its main features are: the creation and management of any number of specialised glossaries/dictionaries in any language; it can handle large files (i.e. over 50K entries); it allows users to have several glossaries open at the same time; and it has a rapid and easily configurable search functionality that can be customised to search for all terms, translation candidates and associated descriptive fields, either in all glossaries or in a specific one. Another interesting feature is the drag and drop functionality, which enables to easily insert words into Microsoft documents, for instance.

Lingo is a simple and user-friendly software that offers an effective way to create and manage multilingual glossaries in any language. Additionally, it permits to manually add an infinite number of customised fields into each entry, such as definitions, URLs, synonyms, antonyms, contextual information, notes or any other desirable field. However, it is platform dependent and does not import from or export to common formats like Microsoft Word or Excel.

[9] www.lexicool.com/soft_lingo2.asp

UniLex[10] is a free terminology management tool created by Acolada GmbH for Windows. It aims to help interpreters and translators prepare, use and manage bilingual glossaries or dictionaries in approximately 30 different languages. UniLex offers a variety of search functions and the possibility to combine user glossaries or dictionaries with a full range of dictionaries available in the UniLex series (e.g. Blaha: Pocket Dictionary of Automobile Technology German/English), which can be acquired as single user versions or as network versions for collaborative purposes. UniLex can also be used in a network environment, which allows users to exchange glossaries or dictionaries. Nevertheless, this additional feature requires a commercial license.

In general, UniLex is not only capable of managing user bilingual glossaries or dictionaries, but also dictionary titles from renowned publishers, which are sold by the company to be consulted within UniLex. However, it only works on Windows and does not handle multilingual glossaries.

TermX[11] is a simple and easy-to-use commercial TMS created by Translex Publishing for Windows. Apart from the usual functionalities that TMS offer (such as add, view, search, edit and remove terminology), TermX permits to add contextual information (relating to the use of the term in a specific context), source information (how and where the term was collected) and up to 6 translation equivalents for each individual source term entry. Similar to Intragloss, this tool also allows the user to associate a term to a domain, which then can be used as a filter to search for terminology in a specific sub-area of knowledge. TermX provides a native format for the management and exchange of terminology, as well as import and export capabilities in the most widely used storage formats, like CSV (Comma Separated Values), plain text, MS Excel, XML, RTF, HTML, MultiTerm and PDF.

In short, TermX aims to help interpreters and translators prepare, use and maintain multilingual glossaries in any language outside of the a Computer-assisted Translation (CAT) environment whilst making all data readily available for import and use in the CAT environment when needed.

[10] www.acolada.de/unilex.htm

[11] www.translex.co.uk/software.html

Terminus[12] is a commercial TMS designed by interpreters for interpreters working on Windows. It permits to organise multilingual terminology (up to 5 languages per glossary) into different subjects. Terminus associates terminology with one or more subjects (i.e. domains). Each term has one main subject and as many additional subjects as the user needs. These descriptors are important as they allow users to search and export specific terminology from these pre-defined subjects. Moreover, when searching for terminology, they can be used to limit the search (e.g. display all the terms stored in a particular subject). Especially on extremely large databases, this may reduce the number of terms that match the search criteria.

Terminus is an easy-to-use flexible tool as it enables the users to classify terms into different subjects, to import terminology lists from plain text and MS Excel files, to export results alphabetically or grouped together by the main subject and sorted within each subject to plain text, RTF and PDF. Another interesting feature is the way that records can be displayed by using different colours for different languages.

Table 1 provides a comparative summary of the main features that characterise the TMS described above. Overall punctuations have been assigned for relevance and wealth of functionalities.

3.2 Web-based TMS

ASPLex[13] is a commercial TMS created by TransLex Publishing and based on MS Access. It can be described as a web-based terminology tool capable of maintaining terminology through an online portal with access rights. In other words, ASPLex easily permits to share glossaries among users within the portal (i.e., authorise who can access, view, edit, export, import and print data). As expected, the platform enables to add, view, edit, remove and search for terminology. When adding a new entry, ASPLex allows users to add contextual information (such as grammar attributes, specialised domain, context, abbreviation or author's notes) and up to 6 translation equivalents for each individual source term entry. Nevertheless, the MS Access database file can be extended to include more descriptive elements. The search within ASPLex can be performed at different levels, e.g. according to domain, source term, change date, etc.

To sum up, ASPLex permits to easily create and manage any number of glossaries in any language, share glossaries with other users, import glossaries from plain text and MS Excel, and export them to MS Word, Excel, plain text, XML and PDF.

Interpreters' Help[14] is a powerful and free TMS designed not only to manage multilingual glossaries but also to manage job assignments and clients. Assignments can be created for both personal and community usage, the last one permits to share assignments privately with other Interpreters' Help members. When sharing an assignment with team members, they can comment on it, view assignment details, view and edit glossaries related to the assignment, download assignments files and view the assignment's client page. Interpreters' Help also has the option to make a glossary publicly available to the Interpreters' Help community. Apart from that, this tool keeps a history of all the assignments, it allows to easily find assignments by client and material that was used for a previous assignment, upload assignment files and attach glossaries. Moreover, it permits to create, edit, search and view glossaries; to add, view, remove and edit entries; to add an unlimited number of translation terms; to add a variety of contextual information (such as comment, category, definition, acronym, amongst others); to easily move or remove columns; to add a glossary to the favourites group; to add and remove tags to and from a glossary; to export a glossary to Excel or PDF; to import from MS Word, Excel, Libreoffice/ Openoffice and CSV; to view a printable version of a glossary, and to copy glossaries, either duplicate our glossaries or copy a public one to our account. Interpreters' Help also has a Mac OS version called Boothmate, which permits to access glossaries offline. This standalone version synchronises with the website and can be used in the booth even without an Internet connection. It is important to mention that BoothMate only allows users to search for terminology, not to edit term entries or glossaries.

Interpreters' Help can be considered one of the most complete TMS freely available on the market. Both versions were designed to be a companion tool not only for users who need to search for glossaries in the booth, but also to those who are looking for a user-friendly and

[12] www.wintringham.ch/cgi/ayawp.pl?T=terminus
[13] www.termnet.nl/ASPLex.html
[14] www.interpretershelp.com/

Feature	Intragloss 1 (2014)	InterpretBank 3.102 (2014)	Intraplex 2.1.1.47 (2012)	SDL MultiTerm 2014 (2013)	AnyLexic 4 (2011)	Lingo 4 (2011)	UniLex 0.9 (2007)	TermX (2013)	Terminus 3.1 (2009)
Manages multiple glossaries ⟨no=0; yes=10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	no ⟨0⟩	yes ⟨10⟩	yes ⟨10⟩
Nº of possible working languages ⟨≤100=4; >100=7; unlimited=10⟩	180 ⟨7⟩	35 ⟨4⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	30 ⟨4⟩	unlimited ⟨10⟩	unlimited ⟨10⟩
Nº of languages per glossary allowed ⟨≤3=5; ≥4=10⟩	2 ⟨5⟩	2 ⟨5⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	2 ⟨5⟩	6 ⟨10⟩	5 ⟨10⟩
Nº of descriptive fields ⟨non=0; 1=3; [2-5]=7; >5=10⟩	4 ⟨7⟩	4 ⟨7⟩	non ⟨0⟩	>5 ⟨10⟩	1 ⟨3⟩	>5 ⟨10⟩	2 ⟨7⟩	> 5⟨10⟩	2 ⟨7⟩
Handles documents ⟨no=0; yes=10⟩	yes ⟨100⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩
Unicode compatibility ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩	yes ⟨5⟩	yes ⟨5⟩
Imports from ⟨1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	MS Word, Excel & Plain Text ⟨3⟩	MS Word, Excel, TMEX & Plain Text ⟨4⟩	MS Word, Excel & Plain Text ⟨3⟩	MS Word, Excel & other CAT formats ⟨5⟩	Excel, Plain Text & AEF ⟨3⟩	TMX & Plain Text ⟨2⟩	Plain Text ⟨1⟩	MS Word, Excel & other CAT formats ⟨5⟩	Excel & Plain Text ⟨2⟩
Exports to ⟨non=0; 1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	MS Word & Excel ⟨2⟩	MS Word, Excel, TMEX, Android & Plain Text ⟨4⟩	MS Word, Excel & Plain Text ⟨3⟩	MS Word, Excel & other CAT formats ⟨5⟩	Excel, Plain Text & AEF ⟨3⟩	TMX & Plain Text ⟨2⟩	Plain Text ⟨1⟩	MS Word, Excel & other CAT formats⟨5⟩	RTF, PDF & Plain Text ⟨3⟩
Embedded online search for translation candidates ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩
Interface's supported languages ⟨1=1; [2-5]=3; >5=5⟩	English ⟨1⟩	English ⟨1⟩	English ⟨1⟩	> 5 ⟨5⟩	> 5 ⟨5⟩	English ⟨1⟩	English + 3 ⟨3⟩	English ⟨1⟩	English ⟨1⟩
Remote Glossary Exchange ⟨no=0; yes=5⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩
Well-documented ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩	yes ⟨5⟩	yes ⟨5⟩
Availability ⟨proprietary without demo=1; proprietary with demo=3; free=5⟩	proprietary without demo ⟨1⟩	proprietary with demo ⟨3⟩	proprietary with demo ⟨3⟩	proprietary without demo ⟨1⟩	proprietary with demo ⟨3⟩	proprietary with demo ⟨3⟩	free ⟨5⟩	proprietary without demo ⟨1⟩	proprietary with demo ⟨3⟩
Operating System(s) ⟨1=1; 2=3; ≥3=5⟩	Mac OS X ⟨1⟩	Windows & Android ⟨3⟩	Windows & iOS ⟨1⟩	Windows ⟨1⟩	Windows ⟨1⟩	Windows ⟨1⟩	Windows ⟨1⟩	Windows ⟨1⟩	Windows ⟨1⟩
Other relevant features ⟨subjective analysis=max. 5⟩	allows to highlight terms in the documents and merge a glossary with a document making it annotated to be printed ⟨5⟩	the MemoryMode helps to memorise bilingual glossaries ⟨4⟩	permits to have several glossaries open at the same time ⟨2⟩	it is a concept oriented-tool and permits to add illustrations into each entry ⟨5⟩	allows to share within a group of AnyLexic users ⟨1⟩	permits to add an unlimited number of descriptive fields ⟨5⟩	-	availability to import and export from and to CAT tools ⟨5⟩	demo version only limits the number of entries ⟨1⟩
Final Mark	67	60	55	77	64	64	27	68	56

Table 1: Comparative standalone TMS: *Intragloss*, *InterpretBank*, *Intraplex*, *SDL MultiTerm*, *AnyLexic*, *Lingo*, *UniLex*, *TermX* and *Terminus*.

straightforward terminology management tool.

Examples of most innovative and consequently more expensive web-based terminology management solutions on the market today are **WebTerm**[15], **Acrolinx**[16], **Termflow**[17] and **flashterm**[18]. Apart from the basic options offered by the aforementioned web-based TMS (e.g. create, edit, view, remove and group terms into domains; add contextual information the each entry; import from and export to e.g. plain text or CSV; manage multilingual glossaries; and, share glossaries with a group of users), these tools offer more sophisticated features, such as:

- extract multilingual terminology from translation memories, PDF, XML, etc. (e.g. Acrolinx and Termflow);
- import and export terminology in industry-standard exchange formats (e.g. OLIF, XML, MTF, TBX, TMX, MARTIF, CSV, SDL's MultiTerm format) (e.g. Acrolinx, WebTerm and Termflow);
- advise whether a translation term is preferred or prohibited in a specific domain (e.g. Acrolinx, Termflow and flasterm);
- integrate a reference database to store client instructions, internal procedures, employee contact information and other useful information to the interpretation or translation service (e.g. LogiTerm and WebTerm)
- an administrative control in which the project manager can select the fields that are displayed, which functions can be used and which settings can be changed (e.g. WebTerm, Acrolinx, Termflow, and flasterm).

Bering this in mind, these tools can be considered more sophisticated than the standalone TMS previously mentioned since they include more advanced features and offer professional support, as they were specially designed for commercial purposes. Although they were not built to help interpreters during the interpretation process, they can be extremely useful before the interpreting service as they allow them to store and share terminology more easily, especially for companies who have a considerable number of employers or for free-lance interpreters in a collaborative environment.

Due to space constraints, only some of all the available web-based TMS on the market can be mentioned. Nevertheless, there are some TMS worth to mention, such as:

- **AcrossTerm**[19] a centralised TMS for the entire company terminology;
- **i-Term**[20] a state-of-the-art terminology and knowledge management tool which allows to store, structure and search online for knowledge about concepts;
- **Multitrans Prism**[21] an innovative client-server software solution that integrates project and business management, translation memory, and terminology management;
- **qTerm**[22] a web-based TMS that permits to identify, define, and translate critical terminology. It also provides a detailed explanation of each term's use, including the context, language, and history of use;
- **TermWiki**[23] a seamless collaborative TMS that aims to collect every term in every subject in the world and make it available in every language. It permits to search for translation candidates.

Table 2 provides a comparative summary of the main features that characterise the web-based TMS mentioned above. As in previous cases, overall punctuations have also been offered to serve as a quick guide or checklist for interpreters.

3.3 Mobile TMS

Mobile terminology applications (or TMS apps) are undoubtedly the next step in this ever-evolving domain of term management. TMS apps are systems which have been developed or optimised for small handheld devices, such as mobile phones, smartphones, iPads or PDAs, among others. Some of the most popular ones are Glossary Assistant and The Interpreter's Wizard.

[15] www.star-group.net/en/products/webterm.html
[16] www.acrolinx.com/platform-services/terminology-management/
[17] www.termflow.de/
[18] www.flashterm.eu/home
[19] www.across.net/en/
[20] www.iterm.dk/
[21] http://linguistech.ca/MultiTrans_EN
[22] http://advancedlanguage.com/services/translation/terminology-management/
[23] www.termwiki.com/

Feature	**ASPLex** (2013)	**Interpreters' Help** beta (2014)	**WebTerm** 6 (2014)	**Acrolinx** (2014)	**Termflow** (2013)	**flashTerm** (2015)
Manages multiple glossaries ⟨no=0; yes=10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩
Nº of possible working languages ⟨≤100=4; >100=7; unlimited=10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	>100 ⟨7⟩
Nº of languages per glossary allowed ⟨≤3=5; ≥4=10⟩	6 ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩	>4 ⟨10⟩	>4 ⟨10⟩
Nº of descriptive fields ⟨non=0; 1=3; [2-5]=7; >5=10⟩	>5 ⟨10⟩	unlimited ⟨10⟩	>5 ⟨10⟩	3 ⟨7⟩	>5 ⟨10⟩	>5 ⟨10⟩
Handles documents ⟨no=0; yes=10⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	yes ⟨10⟩	no ⟨0⟩
Remote Glossary Exchange ⟨no=0; yes=10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩	yes ⟨10⟩
Unicode compatibility ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩
Imports from ⟨1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	MS Word, Excel & other CAT formats ⟨5⟩	MS Word, Excel, Open/Libreoffice & CSV ⟨4⟩	> 5 ⟨5⟩	> 5 ⟨5⟩	> 5 ⟨5⟩	-
Exports to ⟨non=0; 1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	MS Word, Excel & other CAT formats ⟨5⟩	Excel & PDF ⟨2⟩	> 5 ⟨5⟩	> 5 ⟨5⟩	CSV & TBX ⟨2⟩	> 5 ⟨5⟩
Embedded online search for translation candidates ⟨no=0; yes=5⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩	no ⟨0⟩
Interface's supported languages ⟨1=1; [2-5]=3; >5=5⟩	English+3 ⟨3⟩	English ⟨1⟩	> 5 ⟨5⟩	English ⟨1⟩	English & Deutsch ⟨2⟩	English & Deutsch ⟨2⟩
Well-documented ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	yes ⟨5⟩	no ⟨0⟩	yes ⟨5⟩
Availability ⟨proprietary without demo=1; proprietary with demo=3; free=5⟩	proprietary with demo ⟨3⟩	free ⟨5⟩	proprietary without demo ⟨1⟩	proprietary with demo⟨3⟩	proprietary without demo ⟨1⟩	proprietary with demo ⟨3⟩
Other relevant features ⟨subjective analysis=max. 5⟩	-	clean and straightforward TMS that allows to add an unlimited number of translation terms ⟨5⟩	-	simple interface and allows to add illustrations ⟨3⟩	-	clean interface and allows to add illustrations ⟨3⟩
Final Mark	76	77	76	74	75	78

Table 2: Comparative web-based TMS: *ASPLex*, *Interpreters' Help*, *WebTerm*, *Acrolinx*, *Termflow* and *flashterm*.

Glossary Assistant[24] is a user-friendly multilingual glossary management application created by a professional team of interpreters for Android devices. Specially designed to simultaneous/consecutive interpreting, Glossary Assistant allows users to have a comfortable viewing of glossaries on Android-tablets (limited on smartphones). It enables to create, remove and manage multilingual glossaries (glossaries can be maintained up to 10 languages); to add, edit and remove entries from/to a glossary; to search for terms either in a specific language or in all the languages, and to re-arrange and sort columns by language and alphabetically, respectively. The glossaries can be imported and exported from/to Unicode plain text files. In order to import glossaries from third-party application, Glossary Assistant only requires those glossaries to be stored as a tab delimited Unicode text file. Glossary Assistant has a free (4 glossaries maximum, with a maximum of 250 rows per glossary) and a commercial version (which does not has restrictions). A PC version is also available for free without restrictions. In both versions, PC and Android, all the glossaries are internally stored in a database and they can be exchanged between them.

The Interpreter's Wizard[25] is a free iPad application capable of managing bilingual glossaries in a booth. It is a simple, fast and easy-to-use application that helps the interpreter to search and visualise terminology in seconds. The system includes rapid and easily configurable search functionality that can be customised to search for all terms, translation candidates either in all glossaries or in a specific one. Nevertheless, all the imported glossaries need to be previously created and converted online to the proprietary format, and it does not allow users to export glossaries.

Table 3 provides a comparative summary of the main features that characterise these two mobile TMS. As in previous cases, overall punctuations have also been offered to serve as a quick guide or checklist for interpreters.

4 Comparative Analysis

Although the aforementioned Terminology Management Systems (TMS) can be used to prepare a given interpretation of any kind according to the interpreters' requirements identified in section 2, these systems differ from one another in their functionalities, practical issues, degrees of user-friendliness and target audience (i.e. individual or enterprise usage). Therefore, it is necessary to establish a set of specific and measurable features that permit us to assess and distinguish the different tools concerning individual's and company's needs in such a way that the results would be useful for both potential customers as well as to the designers of such systems. Departing from the conclusions drawn from the literature review (see section 1 and 2) and the description of the terminology tools analysed in section 3.1, 3.2 and 3.3 (standalone, web-based and mobile TMS, respectively), this section provides an extensive analysis of these TMS based on our own practical set of measurable features. For instance, the "freedom to define the basic structure" identified by Rodríguez and Schnell (2009) was reformulated into several practical measurable features, such as "**Nº of descriptive fields**", "**Nº of working languages**" and "**Nº of languages per glossary**". Moreover, the possibility of "**developing multilingual mini-databases**", also identified in their study, was reconsidered as measurable features by means of the following criteria: "**Manages multiple glossaries**" and "**Nº of languages per glossary**". Another example is the "**Remote Glossary Exchange**" measurable feature, which was inferred from the study conducted by Bilgen (2009), who identified the need to exchange terminological information.

After a careful analysis of the priorities for the design and features to be included in a terminology management tool reported in Moser-Mercer (1992), Bergenholtz and Tarp (2003), Tarp (2008), Spohr (2009), Rodríguez and Schnell (2009), Bilgen (2011) – see section 2 for more details, we identified 15 main features. Although some of them are pointed out as fundamental due to their extreme importance when assisting interpreters before and during an interpretation service, others are mostly related with the tools' design and surrounding.

In an attempt to standardize these 15 features into a discriminative 0-100 scoring system, we used the EAGLES framework for the evaluation of NLP (Natural Language Processing) systems as a reference to divide these features into two categories: fundamental and secondary. The EAGLES (1996) report includes formalisms of evaluation procedures for various types of systems according to

[24] http://swiss32.com

[25] http://the-interpreters-wizard.appsios.net/

Feature	**Glossary Assistant** 1.2 (2015)	**The Interpreters' Wizard** 2.0 (2011)
Manages multiple glossaries ⟨no=0; yes=10⟩	yes ⟨10⟩	yes ⟨10⟩
Nº of possible working languages ⟨≤100=4; >100=7; unlimited=10⟩	unlimited ⟨10⟩	unlimited ⟨10⟩
Nº of languages per glossary allowed ⟨≤3=5; ≥4=10⟩	10 ⟨10⟩	2 ⟨5⟩
Nº of descriptive fields ⟨non=0; 1=3; [2-5]=7; >5=10⟩	non ⟨0⟩	non ⟨0⟩
Handles documents ⟨no=0; yes=10⟩	no ⟨0⟩	no ⟨0⟩
Unicode compatibility ⟨no=0; yes=5⟩	yes ⟨5⟩	yes ⟨5⟩
Imports from ⟨1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	Plain Text ⟨1⟩	Proprietary Format ⟨1⟩
Exports to ⟨non=0; 1=1; 2=2; 3=3; [4-5]=4; >5=5⟩	Plain Text ⟨1⟩	non ⟨0⟩
Embedded online search for translation candidates ⟨no=0; yes=5⟩	no ⟨0⟩	no ⟨0⟩
Interface's supported languages ⟨1=1; [2-5]=3; >5=5⟩	English ⟨1⟩	English ⟨1⟩
Remote Glossary Exchange ⟨no=0; yes=5⟩	no ⟨0⟩	no ⟨0⟩
Well-documented ⟨no=0; yes=5⟩	yes ⟨5⟩	no ⟨0⟩
Availability ⟨proprietary without demo=1; proprietary with demo=3; free=5⟩	proprietary with demo ⟨3⟩	free ⟨5⟩
Operating System(s) ⟨1=1; 2=3; ≥3=5⟩	Android and Windows ⟨3⟩	iOS (iPad) ⟨1⟩
Other relevant features ⟨subjective analysis=max. 5⟩	user-friendly and intuitive interface ⟨4⟩	quick performance ⟨1⟩
Final Mark	53	39

Table 3: Comparative mobile TMS: *Glossary Assistant* and *The Interpreter's Wizard*.

their general quality characteristics and their definitions: functionality, reliability, usability, efficiency, maintainability and portability. It is important to mention that the EAGLES methodology used as a starting point the ISO 9126 standard for software quality (ISO/IEC, 1991). Although all the reported characteristics are important to any software, in this work our main focus is on the functionality of the software. Thus, we considered fundamental all the features related with the software's functionality ("*A set of attributes that bear on the existence of a set of functions and their specified properties. The functions are those that satisfy stated or implied needs*"), such as the management of multiple glossaries, the number of possible working languages permitted by the tool, how many of these languages can be used at the same time per glossary, the number of descriptive fields allowed per glossary entry and the possibility of managing terminology with preparation documents. The remaining 10 features, which are related with reliability, usability, efficiency, maintainability and portability where categorised as secondary. In detail, the features classified as fundamental to a terminology tool was given 10 points and 5 points to the secondary ones - except for web-based TMS, in which we removed one feature and considered 6 as fundamental and 8 as secondary. Then, these features were used to evaluate the seventeen tools (9 standalone, 6 web-based and 2 mobile) presented in sections 3.1, 3.2 and 3.3 and to assess which one is the most complete, both considering each sub-group separately and all the tools together.

The first feature clarifies whether the tools were designed to handle multiple glossaries in their interfaces at same time (Manages multiple glossaries). The next two features are somehow related. The **Nº of possible working languages** describes how many different working languages are permitted by the application. Then, considering these working languages, how many of them can be used at the same time per glossary (**Nº of languages per glossary allowed**). The next feature is related with all types of descriptive fields that these tools allow to add to each glossary entry (**Nº of descriptive fields**). The possibility of managing terminology with preparation documents (**Handles documents**) is another relevant feature for interpreters seeking for tools capable of highlighting terms in documents, for example. Equally import is the Unicode support (**Unicode compatibility**) as it provides a unique number for every character, no matter what the platform, the program, or the language is. In other words, an application that supports full Unicode means that it has support for any ASCII or non-ASCII language, such as Hebrew or Russian, two non-ASCII languages. **Imports from** and **Exports to**, as its name suggests, represents the supported input and output formats. The **Embedded online search for translation candidates** is a relevant add-in for terminology tools, as it permits to focus the search for terminological candidates within the tool. Despite the fact that all the tools have English as a default language, the support of multiple languages (**Interface's supported languages**) is another important

feature as this would definitely increase the number of potential users that a terminology tool can reach. The **Remote glossary exchange** feature is important when co-operating with other working partners remotely is required, as in collaborative interpreting and crowd-sourcing. The next three features are related with the available documentation, their availability and platform dependency (**Well-documented**, **Availability** and **Operating System(s)**, respectively). Finally, the last row presents some unique characteristics along with some relevant comments (**Other relevant features**).

Based on this comparative analysis, none of the investigated terminology tools exhibit all the desirable features. Nevertheless, SDL MultiTerm, TermX and Intragloss are the best classified standalone TMS with 77, 68 and 67 points out of 100, respectively (see Table 1). This is not surprising because SDL MultiTerm is the most expensive standalone tool nowadays available on the market and, apart from that, it has been developed for more than 20 years. Also developed for commercial purposes, TermX was created by a team of professionals focused on linguistic services, such as translation and terminology management. The score of Intragloss, released in 2014 as a stable version, is neither surprising due to its novelty and design purposes, i.e. it was specifically developed by interpreters for interpreters and, thus, it is entirely tailored to their needs. All three offer a user-friendly interface to easily store, manage and search for multilingual terminology and definitions. On the other hand, UniLex, Intraplex and Terminus got the worst scores due to the lack of features offered (27, 55 and 56, respectively). About the remaining tools (AnyLexic, Lingo, InterpretBank), they have similar features, which resulted in similar scores (64, 64 and 60, respectively). It is worth mentioning that the three best-classified tools were released between 2013 and 2014 and those that got lower scores were released between 2007 and 2012, which means that recent standalone TMS are better designed to assist interpreters.

Sharing terminology is extremely important because it allows users to improve glossaries, making them more uniform, complete and correct across subjects and domains, which can only be accomplished collaboratively. Moreover, sharing terminology is the only way to collect most terms in most subjects in the world and make this knowledge available in every language. Bearing this in mind, web-based TMS take advantage of cutting-edge technologies to fulfil the need for sharing terminology. As we can see in Table 2, all the 6 web-based TMS analysed got similar scores, ranging from 74 (Acrolinx) to 78 (flahterm). This means that they have similar features and should be investigated in more detail by those who are looking for commercial web-based TMS, especially the prices and the technical support provided. It is also important to notice that all these tools have the released date between 2013 and 2015. Apart from the common options offered by traditional TMS, these web-based systems also integrate a Content Management System (CMS). In other words, web-based TMS, not only offer a set of features to manage terminology in a collaborative environment, but also provide procedures to manage the entire workflow, i.e. a CMS that allows users to manage reference databases to store client instructions, internal procedures, employee contact information and amongst other additional information related to the interpretation or translation service.

Despite mobile TMS do not get acceptable scores when compared with standalone and web-based TMS (Glossary Assistant: 53; The Interpreter's Wizard: 39 - see Table 3) and they do not offer the necessary comfort to manage terminology, they still play an important role when a quick search for terminology is required, e.g. while in a booth.

To sum up, web-based programs obtained higher average score compared with standalone and mobile TMS (76, 60 and 46, respectively). These results can be explained by the companies' effort and the cutting-edge technology used during their development. Another fact that contributes to the increasing interest in web-based TMS is that nowadays companies are more orientated towards developing centralised systems in order to provide uniform services to both staff and clients. Nevertheless, this effort requires higher investment in equipment and manpower to maintain these systems and consequently make them more expensive compared with standalone or mobile TMS. The only exception is the Interpreters' Help tool (still in beta).

5 Conclusions

This paper presents a comprehensive and up-to-date review of the currently available TMS on the market as well as an overview of the most relevant features that these tools should have in order to help interpreters before and during the

interpretation process. Seventeen terminology tools have been described and compared with the aim of assessing them on the basis of a set of 15 features previously identified and a scoring system. This comparative analysis aims at highlighting some of the features that interpreters can expect from the terminology management tools currently available on the market. In addition, the results obtained could guide interpreters when choosing specific tools for a given interpretation project, i.e. the TMS(s) that would best cater for their specific needs, in order to help them work more efficiently, store and share terminology more easily, as well as save time when a looking for a specific feature most suited to a specific interpreting service.

Sharing terminology is extremely important because it allows users to improve terminology by enhancing term coverage and consistence within and throughout domains in a collaborative fashion. Although most of the analysed TMS could be considered to be very flexible when searching for terminology within glossaries and that they can help interpreters carry out their terminology management, it appears that none of them can fulfil all interpreters' needs. It is worth mentioning that some tools require a steep learning curve (e.g. Lingo) while others imply a significant financial investment (e.g. SDL MultiTerm, ASPLex, WebTerm and flashterm). Moreover, some tools are fairly basic and more orientated towards creating and managing bilingual or multilingual glossaries rather than more comprehensive terminology records with supporting information (e.g. UniLex and The Interpreter's Wizard).

Our main findings suggest that most TMS are not envisaged to be used by interpreters. Therefore, TMS do not fulfil completely the needs of this group of end-users as regards speed of consultation, intuitive navigation, possibility of updating the terminology record in the interpretation booth, freedom to define the basic structure, multiple ways of filtering data and sharing information, etc. Conversely, those tools devoted to interpreters (and mainly developed by interpreters) are fairly basic and only include a limited number of features. Another important observation is that the most comprehensive, user-friendly and successfully evaluated systems are standalone TMS, which are also greater in number (if considering purely TMS). This fact reinforces the idea that most TMS are not addressed to interpreters as their final users, but rather to translators. Interpreters need the information and terminology gathered during the preparation phase in the interpretation service and it is not always possible to use standalone versions. On the other hand, web-based TMS are more recent and have been created with cutting-edge technology, which may result in standalone TMS losing the race to web-based in the short run. Interestingly enough, mobile TMS are not performing as well as the others. This seems to be in contradiction to the extensive usage of apps and the requirement of accessing websites from a number of different devices these days. Mobile TMS should be seen as a portable interface/middleware to a web-based or standalone TMS, especially suitable for quick terminology searches, although it should also be acknowledged that mobile TMS are still far away from offering the same degree of comfort to manage terminology and/or web-based content.

Given that quality terminology management is a top priority for interpreters, there seems to be a pressing need to design terminology management tools tailored specifically to assist interpreters both prior and during their interpreting services. In this vein, it would be necessary to ascertain interpreters' terminology needs (as opposed to translators'), and then, devote more efforts to the development of web-based and, particularly, mobile TMS in order to provide on-site consultation of glossaries, terminologies, lists of proper names and conversion figures, etc. No doubt, technology-assisted interpreting will offer a challenging and fruitful research niche for many years to come. We are just at the beginning of this long and winding road...

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n. 317471. The research reported in this work has also been partially carried out in the framework of the Educational Innovation Project NOVATIC (PIE 15-145, 2015-2017); the R&D project INTELITERM (ref. n. FFI2012-38881, 2012-2015); the R&D LATEST (Ref: 327197-FP7-PEOPLE-2012-IEF) project; and the R&D Project for Excellence TERMITUR (ref. n. HUM2754, 2014- 2017).

References

Berber, Diana. 2010. *ICT (Information and Communication Technologies) in Conference Interpreting: a survey of their usage in professional and educational settings*. PhD Thesis, University of Turku and Abo Akademi University, Finland and Universitat Rovira I Virgili, Tarragona, Tarragona, Spain.

Bergenholtz, Henning and Sven Tarp. 2003. Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. *Hermes, Journal of Linguistics*, 31:171–196.

Bilgen, Baris. 2011. *Investigating Terminology Management for Conference Interpreters: A User-oriented Study*. LAP Lambert Academic Publishing.

Blasco Mayor, María J. 2005. El reto de formar intérpretes en el siglo XXI. *Translation Journal*, 9(1).

Braun, Sabine. 2006. Multimedia communication technologies and their impact on interpreting. In H. Gerzymisch-Arbogast M. Carroll and S. Nauert, editors, *Audiovisual Translation Scenarios (MuTra'06)*, Copenhagen, Denmark.

Costa, Hernani, Gloria Corpas Pastor, and Isabel Durán Muñoz. 2014a. A comparative User Evaluation of Terminology Management Tools for Interpreters. In *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, pages 68–76, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Costa, Hernani, Gloria Corpas Pastor, and Isabel Durán Muñoz. 2014b. Technology-assisted Interpreting. *MultiLingual 143*, 25(3):27–32, April/May.

Costa, Hernani, Gloria Corpas Pastor, and Isabel Durán Muñoz. 2015. An Interpreters' Guide to Selecting Terminology Management Tools. In *NATO Conference on Terminology Management*, Brussels, Belgium, November.

Costa, Hernani, Anna Zaretskaya, Gloria Corpas Pastor, and Miriam Seghiri. 2016. Nine terminology extraction Tools: Are they useful for translators? *MultiLingual #159*, 27(3), April/May.

de Manuel Jerez, Jesús. 2003. *Nuevas tecnologías y formación de intérpretes*. Editorial Atrio, Granada, Spain.

Durán Muñoz, Isabel. 2012. Meeting Translators' Needs: Translation-oriented Terminological Management and Applications. *The Journal of Specialised Translation*, 18:77–92. Available at: `http://www.jostrans.org/issue18/art_duran.pdf` (Accessed 30 June 2014).

EAGLES. 1996. Evaluation of Natural Language Processing Systems. Technical report, EAGLES Document EAG-EWG-PR.2, October. `http://www.issco.unige.ch/en/research/projects/ewg96/ewg96.html`.

Fantinuoli, Claudio, 2016. *Computer-assisted interpreting: challenges and future perspectives*, chapter Trends in e-tools and resources for translators and interpreters. Brill.

Gile, Daniel. 1987. La terminotique en interprétation de conférence: un potentiel á exploiter. *Meta: Translators' Journal*, 32(2):164–169, June.

Gran, Laura, Angela Carabelli, and Raffaela Merlini. 2002. Computer-assisted interpreter training. *Interpreting in the 21st Century: Challenges and opportunities*, 43(1):277–294.

Kalina, Sylvia. 2010. New Technologies in Conference Interpreting. In *Am Schnittpunkt von Philologie und Translationswissenschaft. Festschrift zu Ehren von Martin Forstner*, pages 79–96. Bern, Peter Lang.

Kelly, Nataly. 2009. Moving toward machine interpretation. *TCWorld*.

Moser-Mercer, Barbara. 1992. Banking on Terminology: Conference Interpreters in the Electronic Age. *Meta: Translators' Journal*, 37(3):507–522, September.

Pöchhacker, Franz. 2007. *Introducing Interpreting Studies*. London and New York: Routledge, 2nd edition.

Rodríguez, Nadia and Bettina Schnell. 2009. A Look at Terminology Adapted to the Requirements of Interpretation. *Language Update*, 6(1):21–27.

Sandrelli, Annalisa and Jesús de Manuel Jerez. 2007. The Impact of Information and Communication Technology on Interpreter Training. *The Interpreter and Translator Trainer*, 1(2):269–303.

Spohr, Dennis. 2009. Towards a Multifunctional Electronic Dictionary Using a Metamodel of User Needs. In *eLexicography in the 21st century: New challenges, new*

applications, Louvain-La-Neuve, Belgium. Presses Universitaires de Louvain.

Tarp, Sven. 2008. *Lexicography in the Borderland Between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Lexicographica: Series maior. Walter de Gruyter, 1^{st} edition.

Tripepi Winteringham, Sarah. 2010. The usefulness of ICTs in interpreting practice. *The Interpreters' Newsletter*, 15:87–99.

Will, Martin. 2007. Terminology Work for Simultaneous Interpreters in LSP Conferences: Model and Method. In Heidrun Gerzymisch-Arbogast and Gerhard Budin, editors, *Proc. of the Marie Curie Euroconferences MuTra: LSP Translation Scenario*, EU-High-Level Scientific Conference Series, Vienna, Austria.

Zhang, Xiaojun, 2016. *Technical Issues on Videoconference-based Interpreting*, chapter Trends in e-tools and resources for translators and interpreters. Brill.